Protection des données de santé lors de leur réutilisation à des fins de recherche au sein des entrepôts de données biomédicaux (EDBM)

Health data protection for secondary use in clinical data warehouse

Christine Riou[1] **Marc Cuggia**[2]

[1] *Département d'Information Médicale, CHU de Rennes, France*
[2] *INSERM LTSI - Equipe Projet Données massives en santé*
Faculté de Médecine de Rennes, France

Résumé

Introduction : La mise en œuvre des entrepôts de données biomédicaux (EDBM) pose des questions d'ordre juridique, déontologique et éthique. L'objectif de cet article est d'effectuer un état des lieux sur ces points et de formuler des propositions organisationnelles pour l'exploitation des EDBM. Méthodes : les auteurs ont effectué une revue de la littérature dans Pubmed, Google Scholar, Google, Science Direct. Elle a concerné des articles scientifiques, des textes de loi ou réglementaires ainsi que des procédures de mise en œuvre des systèmes publiés. Résultats : l'utilisation des données à caractère personnel pour la recherche en santé est conditionnée à l'accord des patients et à l'autorisation d'autorités de contrôle. Le patient dispose d'un droit d'information et d'un droit d'opposition. Les solutions proposées pour assurer la protection des données dans les EDBM s'appuient essentiellement sur la dé-identification des données et l'organisation des accès. Conclusion : l'information du patient reste délicate. Les procédures de dé-identification sont à améliorer et à certifier. Le chaînage des données s'avère nécessaire pour le croisement avec d'autres sources externes. L'organisation proposée par les auteurs pour l'exploitation des EDBM repose sur la mise en place de structures dédiées ainsi que sur la rédaction de documents cadre et sur la définition de procédures.

Abstract

Background : Implementing clinical data warehouse (CDW) brings out legal and ethical issues. Methods : the authors make a review of recent litterature on these points in Pubmed, Google scholar, Google, Science direct. Il includes scientific papers, texts of laws, and operational procedures. Then they propose an organisation for CDW implementation. Results : use of health personal data required patient's consent and an authorization of competent authorithies. Rights of information and opposition are due to the patient. To assure data protection in CDW measures rely on data deidentification and hierarchical management of access. Conclusion : patient information remains difficult. Deidentification procedures are to be improved and

Articles longs des 15es Journées francophones d'informatique médicale, JFIM 2014, pages 162–187
Fès, Maroc, 12–13 juin 2014

certified for french language. Crossing data needs chaining them. The organisation proposed by the authors lies on specific structures and documents (policy for the use of CDW, patient information note, charter for good use).

Mots clés : Confidentialité, dé-identification, entrepôt de données biomédicales, dossier patient informatisé, recherche en santé

Keywords : *Confidentiality, deidentification, data wharehouse, electronic patient record, secondary use, medical research*

1 Introduction

Avec l'informatisation des établissements de santé, la dématérialisation des données patient modifie le rapport à l'information médicale. Ces données sensibles, hétérogènes et multidomaines deviennent à présent massives (notion de bigdata) et potentiellement accessibles de façon ubiquitaire. Ces données constituent un gisement d'informations majeur qui peut être réutilisé pour la recherche. L'exploitation concomitante des données cliniques et des données issues du génome (données OMICS) et plus généralement la recherche de biomarqueurs in silico répond aux besoins actuels de la médecine translationnelle et personnalisée.

Pour cela, les établissements se dotent de nouveaux outils appelés entrepôts de données biomédicaux (EDBM) permettant la fouille et l'exploitation de ces données. Cependant, leur mise en oeuvre pose un certain nombre de questions, tant sur le plan de la sécurité des données, que sur les plans réglementaire, déontologique et éthique. Ainsi, selon la loi Informatique et libertés [1], l'accès aux données à caractère personnel suppose que le patient ait donné son accord[1] pour une étude précise mobilisant des données bien définies[2]. Or, les EDBM contiennent par définition l'exhaustivité des données patient rétrospectives d'un établissement. Il est donc impossible de connaître à l'avance précisément l'objet des études qui seront réalisées grâce à l'entrepôt et la nature des données qui seront utilisées. La loi définit un principe de parcimonie consistant à n'utiliser strictement que les données nécessaires et suffisantes à l'étude. Or dans un EDBM, le principe de fouille et de recherche d'information suppose l'accès et le traitement d'informations beaucoup plus larges. La loi actuelle est facilitatrice pour les professionnels de santé qui exploitent les données des patients dont ils ont eu la charge[3]. Or, un EDBM permet une exploitation transversale des données par des utilisateurs qui n'ont pas été nécessairement impliqués dans la prise en charge du patient. Ce cas de figure peut être sujet à interprétation, simple déclaration ou autorisation « recherche médicale »[4], plus contraignante à mettre en œuvre et longue à obtenir.

D'autres questions se posent : Quel est le rôle du patient dans ce dispositif ? Comment assurer la protection de sa vie privée ? Comment organiser les accès à l'entrepôt ?

L'objectif de ce travail est d'effectuer un état des lieux des solutions adoptées pour assurer la

[1] sauf dérogation spécifique de la Commission Nationale Informatique et Libertés (CNIL)

[2] Article 30 de la loi informatique et libertés

[3] Article 53 de la loi informatique et libertés

[4] Chapitre IX de la loi Informatique et libertés

protection des données de santé lors de leur réutilisation pour la recherche dans les entrepôts de donnés biomédicaux.

A partir de cette analyse les auteurs formulent des propositions opérationnelles dans le respect de la réglementation française et adaptées au cadre organisationnel des établissements de santé permettant l'exploitation des documents structurés et textuels dans les EDBM tout en garantissant la protection des données des patients.

2 Etat de l'art

Nous commencerons par définir les concepts clefs abordés dans cet article.

2.1 Les entrepôts de données biomédicaux

Les entrepôts de données biomédicaux ont vu le jour aux Etats-Unis au milieu des années 2000 [2-3], ils sont en cours d'implémentation dans les autres pays notamment en Europe [4]. Un entrepôt de données biomédical (EDBM) regroupe les données patient, produites lors de leurs prise en charge, stockées dans les différentes bases de données du système d'information de l'établissement (dossier patient informatisé, système de gestion des laboratoires, système d'information de radiologie, système de dispensation des médicaments, système de recueil d'activité des blocs opératoires et plateaux médico-techniques). Il peut-être étendu aux données médico-économiques mais aussi aux données OMICS.

Il permet une indépendance par rapport au système de production et une centralisation des données. Il intègre des outils de fouille de données, autorisant des requêtes croisées sur des données structurées ou textuelles, sur des critères biologiques et cliniques par exemple, ainsi que des outils de visualisation (timeline, cluster, regroupement géographique) [4].

Il peut être utilisé dans différents domaines :

- Recherche clinique [5] : étude de faisabilité, recherche de patients éligibles (Pre-screening) à un protocole de recherche, alimentation des cahiers d'observation électronique (e-CRF)
- Epidémiologie : constitution de cohortes épidémiologiques, alimentation de registres [6]
- Evaluation des pratiques professionnelles
- Vigilances : la détection d'effets indésirables [7] ou la détection des infections nosocomiales [8]
- Etude médico-économique [9]
- Recherche translationnelle : [10]

Leur exploitation est centrée dans une approche populationnelle. A titre d'exemple nous présentons 3 outils actuellement opérationnels :

- Informatics for Integrating Biology and the Bedside (i2b2) [3]

  Il s'agit d'un système développé à la faculté de médecine de Harvard et implanté dans plusieurs centres hospitalo-universitaires américains et européens [11-12]. I2B2 est un outil de recherche translationnelle permettant l'analyse combinée de données cliniques et génomiques. Cette plateforme Open Source connaît une large adoption par le monde

académique et industriel.

- Standford Translationnal Research Integrated Database Environment (STRIDE) [2] développé et mis en œuvre à l'université de Stanford.

Dans l'entrepôt STRIDE il est mis à disposition des chercheurs de l'université de Stanford trois outils [13] :

- Outil pour étude de faisabilité avec production de résultats agrégés (cohort discovery tool),
- Outil pour la constitution de cohorte (cohort data review tool)
- Outil d'extraction de données (data extraction tool).

- Roogle entrepôt déployé en France dans différents établissements de santé [14].

  Le système Roogle est un moteur de recherche développé par l'université de Rennes 1 et le CHU de Rennes. Il permet la réexploitation combinée des données structurées et textuelles, la réalisation de requêtes centrées sur une population, et également l'interrogation au sein d'un dossier patient individuel.

2.2 Les données à caractère personnel

Un entrepôt de données biomédical contient des données de santé à caractère personnel. On entend par donnée à caractère personnel « toute information relative à une personne physique identifiée ou qui peut être identifiée directement ou indirectement par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres ». [1].

Les conditions de protection de ces données lors de leur traitement sont fixées par la loi, comme la loi Informatique et libertés en France [1], l'Health Insurance Portability and Accountability Act (HIPAA) [15] ou encore la Common Rule aux Etats-Unis, the Data Protection Act au Royaume Uni [16].

Dans tous les pays, l'utilisation de données de santé à caractère personnel pour un projet de recherche est conditionnée à l'accord des patients et à l'autorisation d'autorités de contrôle garantissant ainsi la protection des personnes, la confidentialité, le bien fondé et la sécurité du traitement. L'autorisation ainsi que l'accord du patient sont donnés pour chaque traitement. En France les comités de protection des personnes (CPP) jouent un rôle éthique, le Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS) apporte un avis méthodologique et la Commission nationale informatique et libertés (CNIL) s'assure que le traitement est effectué conformément à la législation [17]. Aux Etats Unis les Intitutional Review Boards (IRB) mis en place dans les différentes institutions autorisent les traitements après s'être assurés que le projet ne pose pas de problème éthique et que les droits des personnes participant au projet de recherche sont respectés [18]. Il en est de même au Canada pour les Research Ethic Boards (REB) et au Royaume Uni pour les Research Ethics Committees (RECs). Au Royaume Uni chaque institution nomme une personne, le Caldicott Guardian, responsable de la protection des données, garant du respect de la législation en matière de confidentialité notamment lors du partage d'informations [19]. Celle-ci est consultée lors de la réutilisation de données de santé.

Les traitements concernant des données anonymes ne sont pas concernés par les lois.

2.3 Les données anonymes

En France au sens de la CNIL [20], les données sont considérées anonymes si elles ne permettent pas d'identifier, directement ou indirectement par regroupement, une personne physique. Au niveau européen, les principes édictés dans la directive 95/46/CE relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel [21] ne s'appliquent pas « aux données rendues anonymes d'une manière telle que la personne concernée n'est plus identifiable ».

Il n'y pas de procédé permettant d'assurer un anonymat strict des données individuelles. Le terme de dé-identification est apparu aux Etats Unis après définition dans la loi (HIPPA) des catégories de données identifiantes dénommées protected health information (PHI).

2.4 Données identifiantes

Aux Etats-Unis, la loi, l'HIPAA, a défini 18 catégories de donnés identifiantes, protected health information (PHI) [15]. Elles comportent les noms propres (patients, médecins, établissements de santé…), les adresses et codes de résidence précis, les dates, les numéros attachés au patient (numéro sécurité sociale, numéro assurance, numéro de dossier médical, ….), numéros de téléphone, fax, adresses email, données biométriques, photographies d'identité, les caractéristiques uniques du patient.

L'HIPAA définit trois contextes de dé-identification [22].

- Safe Harbor Standard où les 18 catégories de données identifiantes sont occultées
- Limited Data Set standard (LDS) où les dates et les codes géographiques sont autorisés.
- Statistical standard consiste à certifier que la probabilité d'identifier le patient est très faible.

Au Royaume Uni, dans l'objectif de développer un service pour la réutilisation des données de soins [23], des préconisations ont été définies pour la dé-identification des données à travers le Pseudonymisation Implementation Project (PIP) [24]. Sont considérées comme données identifiantes les noms, adresses, codes postaux, dates de naissance et de décès, numéros d'identification patient (N° NHS, N° local, N° hospitalier, …) et catégorie ethnique [25].

En France la CNIL n'a pas à ce jour établi de liste de données identifiantes. Dans le cadre du Programme de Médicalisation du système d'Information [26], un certain nombre d'informations sont considérées comme sensibles, les dates, le numéro de séjour, l'unité médicale, le code postal.

2.5 Dé-identification des données.

Dé-identification n'est pas synonyme d'anonymat. Il s'agit d'un procédé par lequel les données identifiantes sont supprimées ou occultées. La probabilité de retour à l'identité du patient est alors extrêmement faible. Il persiste cependant un risque de réidentification [27]. Le croisement de fichiers augmente ce risque.

Les préconisations de la CNIL en matière d'anonymisation de fichier s'apparentent à une technique de dé-identification, outre la suppression des données nominatives et des numéros identifiants, remplacement des dates par l'année, des codes géographiques par le département

[28]. Une réflexion est en cours dans le cadre de l'open data sur de nouvelles solutions techniques d'anonymisation.

Nous préférerons utiliser par la suite le terme de dé-identification.

2.6 Droits des patients lors d'un traitement de données à caractère personnel

Un traitement de données à caractère personnel doit se faire dans le respect des droits des patients. Conformément à la loi informatique et libertés le patient dispose d'un droit d'information et d'un droit d'opposition. Les informations contenues dans son dossier médical sont couvertes par le secret. Cependant une dérogation au secret médical pour la recherche est prévue dans la loi conditionnellement aux dispositions inscrites dans le chapitre IX Traitement de données à caractère personnel ayant pour fin la recherche en santé.

Les informations fournies au patient selon l'article 57 de la loi concernent la nature des informations transmises, la finalité du traitement de données, les personnes physiques ou morales destinataires des données, le droit d'accès et de rectification, le droit d'opposition.

L'utilisation de données génétiques est soumise au consentement écrit du patient.

La réutilisation des données de soins implique le plus souvent l'accès aux données nominatives. Selon la loi l'accord du patient est donc requis. Devant la difficulté à retrouver les patients, une dérogation à l'information individuelle[5] peut cependant être obtenue auprès de la CNIL pour une étude donnée.

Le dossier médical contient des informations relevant de l'intimité du patient. Le code de santé publique [29] stipule que « toute personne prise en charge dans un établissement de santé a droit au respect de sa vie privée et au secret des informations le concernant ».

3 Matériel et méthodes

Une revue systématique de la littérature a été effectuée. La recherche documentaire s'est basée sur l'interrogation de Pubmed, Google Scholar, Google, Science direct. Elle concernait des articles scientifiques d'intérêt, et des textes de loi ou réglementaires, ainsi que des procédures de mise en œuvre des systèmes publiés.

Les requêtes ont utilisé les mots clefs suivants : electronic health record (EHR), electronic medical record (EMR), electronic patient record, medical record system, data wharehouse, secondary use, reuse, de-identification, pseudonymisation, patient identifier, anonymisation, anonymized data, privacy, confidentiality, informed consent, consent, clinical research, translationnal medicine.

Nous avons également inclus des publications référencées dans les papiers sélectionnés.

Enfin nous nous sommes appuyés sur le guide de bonnes pratiques pour assurer la confidentialité des données de soins lors de leur réutilisation pour des traitements biostatistiques rédigé par un groupe de travail constitué de chercheurs, d'enseignants de biostatistiques , de médecins de département d'information médicale et de juristes [30].

---

[5] Article 57 loi Informatique et libertés

4 Résultats

Au total, 66 références ont été sélectionnées et utilisées dans notre étude. Elles concernent 40 publications scientifiques, 7 références de texte de loi ou réglementaire, 12 rapports ou guides de référence, 5 site web de projets.

Nous dresserons dans un premier temps un panorama des méthodes déployées pour garantir la confidentialité des informations de santé dans les EDBM. Nous traiterons ensuite un sujet qui lui est lié, celui de l'information du patient.

4.1 Protection des données

Les solutions proposées pour assurer la protection des données dans les entrepôts biomédicaux mettent en avant deux points essentiels, les procédures de dé-identification des données issues des dossiers médicaux et l'organisation des accès.

Nous dresserons dans un premier temps un panorama des méthodes déployées pour garantir la confidentialité des informations de santé dans les EDBM. Nous traiterons ensuite un sujet qui lui est lié, celui de l'information du patient.

4.1.1 Les procédures de dé-identification

Les procédures de dé-identification comportent deux étapes : le repérage puis le traitement des données identifiantes (suppression ou remplacement).

Le repérage

Il est relativement aisé lorsqu'il s'agit de données structurées puisque les algorithmes peuvent s'appuyer sur les éléments et les types de données (par exemple les informations structurées issues du Système d'Information Hospitalier (SIH) comme le nom, l'adresse ou le numéro de séjour).

Il n'en va pas de même pour les documents textuels. Pour ce type de document les algorithmes reposent sur deux types de méthodes [31,32]. Les méthodes symboliques (pattern matching) utilisent des listes de termes (listes de noms, de prénoms, de villes, etc.), des listes de déclencheurs (déclencheurs de personnes : « M., Mme, Melle, Dr, Pr », etc.) et des dictionnaires (langue générale ou langue de spécialité). Les méthodes d'apprentissage automatique nécessitent auparavant l'annotation de textes. Des méthodes hybrides combinent les deux types d'approche.

Pour le traitement de données structurées, il faut citer une autre méthode, la méthode du k anonymat qui permet de repérer les cas identifiables du fait de leur caractère d'unicité et de les affecter de caractéristiques plus génériques [33].

Même si ces méthodes ont démontré de bonnes performances [34, 35], elles présentent cependant des limites [34,36]. L'ambiguïté de certains termes, l'absence de termes dans les dictionnaires ou des fautes d'orthographe peuvent conduire à un mauvais filtrage des données. Ces algorithmes sont souvent développés à partir d'échantillons de compte rendus spécifiques à une institution ou à un domaine, la performance peut ne pas être reproductible. Enfin, la lisibilité et la compréhensibilité de textes dé-identifiés peuvent également être affectées.

Le traitement

Une fois repérées, les données identifiantes sont traitées pour être transformées ou supprimées. Dans ce dernier cas il n'est pas possible de chaîner les données relatives à un même patient.

Certaines données peuvent être floutées, comme les dates remplacées par l'année ou un intervalle de valeurs pour l'âge, ou les codes postaux regroupés pour atteindre un nombre d'habitants suffisant ou remplacés par le département. Elles peuvent aussi être remplacées par des valeurs non signifiantes ne permettant pas de déduire l'identité des patients, on parle alors de pseudonymisation [24, 32, 37]. Les pseudonymes relatifs aux numéros de séjour ou de patient sont générés de telle sorte que toutes les données propres à un séjour ou un patient puissent être regroupées et chaînées entre elles. Un pseudonyme doit être unique et réplicable pour une même donnée à dé-identifier. Un pseudonyme peut être généré soit par une fonction de hachage irréversible, soit par une méthode de cryptage réversible. Le pseudonyme généré n'est alors pas intelligible par un humain. Cette approche est utilisée l'entrepôt de données I2B2 de l'HEGP [11]. En cas de cryptage réversible, le déchiffrement permet de restituer les données initiales en clair. La réversibilité de la pseudonymisation peut également être assurée par une table de correspondance. Pour une meilleure lisibilité des données textuelles, les données identifiantes (comme le nom et le prénom) peuvent aussi être substituées par des informations de même nature puisées aléatoirement dans des tables disponibles. Les dates sont remplacées de telle sorte que la temporalité des évènements soit conservée [32, 33].

Les outils de repérage et de traitement

Plusieurs algorithmes ont été publiés pour la dé-identification de textes en langue anglaise [38, 39]. Ils s'adaptent mal à la langue française. Des outils propres au français ont été développés et évalués comme FASDIM par E Chazard [40] et Medina par C. Grouin ou encore Wapiti [32, 41].

Concernant la pseudonymisation, on peut citer trois services utilisés à l'échelle nationale pour assurer à la fois la pseudonymisation et le chaînage des données patients au delà d'un établissement.

Au Royaume Uni dans le cadre du Pseudonymisation Implementation Project, des outils de dé-identification ont été publiés [24]. Ils sont mis en œuvre dans des projets d'institutions affiliées au NHS [42]. Par ailleurs, l'Agence sanitaire nationale, Health and social care information center (HSIC) [43], met en place un service sécurisé d'appariement des données pour permettre le chaînage des données de santé hospitalières et ambulatoires [44].

En Allemagne un service de pseudonymisation est mis à disposition des organismes de recherche sur une plate-forme gérée par une structure indépendante centrale jouant le rôle de tiers de confiance, the german technology and method platform for networked medical research (TMF) [37]. Il permet d'obtenir un numéro patient unique non signifiant. Il est mis en oeuvre dans le projet d'implémentation de l'entrepôt i2b2 à l'université d'Erlangen [12].

En France des fonctions d'occultation des informations nominatives (FOIN) sont utilisées pour l'anonymisation et le chaînage des séjours des patients des fichiers PMSI (Programme de Médicalisation des Systèmes d'Information) ainsi que pour l'anonymisation (plus exactement la pseudonymisation) des données de la base SNIIRAM (Système National Inter régime de l'Assurance Maladie) qui comprend les données de remboursement de l'ensemble des assurés [45].

4.1.2 Les procédures d'accès et d'exploitation des entrepôts de données biomédicaux

L'exploitation sécurisée d'un entrepôt de données suppose la mise en œuvre d'une procédure d'accès basée sur la définition de droits utilisateur et la traçabilité des accès.

En fonction des cas d'utilisation, par exemple étude de faisabilité, pré-screening ou monitoring, les utilisateurs accèdent respectivement à des données agrégées, dé-identifié ou nominatives.

Ainsi, à chaque niveau de dé-identification correspondent un profil d'utilisateur et un contexte d'utilisation.

Nous présentons ici trois exemples de procédures d'accès et d'exploitation d'entrepôts de données biomédicaux.

I2B2 :

Les droits d'accès à l'entrepôt sont basés sur le niveau de confidentialité des données et le degré de confiance envers l'utilisateur. Le niveau de confidentialité est lié au niveau de dé-identification. L'algorithme de dé-identification des données dans l'entrepôt doit être validé par l'IRB. L'exploitation de l'entrepôt I2B2 aux Etats Unis fait transparaitre cinq niveaux d'accès correspondant à des cas d'utilisation différents [46].

Les niveaux 1 et 2 correspondent à l'accès à des données agrégées anonymes. Ces niveaux sont affectés à des utilisateurs non formés à la sécurité informatique. L'entrepôt est utilisable à partir de postes de travail dont la sécurisation n'est pas optimale.

Concernant le niveau 1, les résultats sont floutés. Ils sont produits à partir d'un entrepôt de données non dé-identifié. Des croisements d'informations sont possibles mais limités et prédéfinis.

Dans le niveau 2, ce sont les effectifs réels qui sont présentés mais calculés à partir d'un entrepôt de données dé-identifié. La fonction de floutage n'est pas dans cas considérée nécessaire.

Ces deux premiers niveaux sont typiquement utilisés pour la réalisation d'études de faisabilité.

Le niveau 3 permet l'accès à des données individuelles structurées de type LDS, l'accès est réservé à des utilisateurs de confiance.

Le niveau 4 autorise le chercheur à visualiser les données individuelles textuelles dé-identifiées de type LDS. L'IRB doit se prononcer sur la confidentialité des données textuelles ainsi dé-identifiées. Là-aussi, l'accès n'est possible que pour des utilisateurs de confiance.

Ces deux niveaux sont utilisés pour la réalisation d'études observationnelles rétrospectives.

Le niveau 5 donne l'accès aux données identifiantes notamment lorsqu'il est nécessaire de recontacter le patient pour par exemple, participer à un essai clinique.

Il est restreint à certains utilisateurs qualifiés et autorisés par l'IRB. Lorsque dans l'entrepôt les données identifiantes (PHI) sont cryptées, seuls les utilisateurs autorisés sont en mesure de décrypter les données.

Pour les accès de niveaux 3 à 5, pour chaque projet de recherche un sous-entrepôt est créé. Il s'agit d'une partition de l'entrepôt global réalisé à partir d'une requête traduisant les critères de sélection. Il ne contient que les dossiers des patients répondant aux critères de l'étude.

L'accès aux données individuelles ne concerne que le sous-entrepôt. Auparavant le projet doit

avoir reçu l'approbation de l'IRB, et en cas d'accès aux données identifiantes une dérogation à l'obtention du consentement du patient doit avoir été délivrée.

Par ailleurs, les chercheurs reçoivent une formation à la sécurité informatique dans le domaine de la santé. Une validation de cette formation est nécessaire pour les accès du cinquième niveau.

STRIDE :

A Stanford l'entrepôt STRIDE [2] est sous la responsabilité d'une structure universitaire (Stanford Center for Clinical Informatics SCCI). Cette dernière détermine et contrôle les accès. Une attestation de formation au droit numérique est nécessaire à l'utilisation de l'entrepôt par les chercheurs.

Aux trois outils d'exploitation des données de l'entrepôt Stride correspondent des niveaux de confidentialité et des conditions d'accès graduées :

L'outil pour étude de faisabilité produit des résultats agrégés sans identification possible des patients (cohort discovery tool). Il est accessible pour les chercheurs de Stanford sur simple demande auprès de la structure en charge de l'entrepôt STRIDE. Les requêtes réalisées avec le « cohort discovery tool » sont tracées et auditées.

Pour utiliser l'outil de constitution de cohortes (cohort data review tool), le chercheur doit disposer de l'autorisation d'un responsable universitaire chargé de la protection des données (school of medicine privacy officer). Les informations consultées sont celles strictement nécessaires pour le projet de recherche. Un premier cadre d'utilisation correspond à un travail préparatoire à la mise en place d'un protocole de recherche, il ne nécessite pas l'accord d'un IRB, seules des données dé-identifiées sont accessibles. Un deuxième cadre d'utilisation est celui d'un projet de recherche déjà validé par un IRB. Ce dernier délivre une autorisation à l'accès aux données individuelles dé-identifiées ou non en fonction du type de projet.

Il en est de même concernant l'outil d'extraction de données (data extraction tool) [13]. Dans l'éventualité où des données identifiantes sont communiquées, le projet doit avoir reçu de l'IRB une dérogation à l'obtention du consentement du patient.

Roogle :

Dans le cadre de l'entrepôt Roogle [14] mis en oeuvre au CHU de Rennes, des règles pour la gestion des accès ont été établies en fonction du type d'étude réalisée sur l'entrepôt [47]. Il est utilisé par les professionnels de santé du CHU ainsi que par les acteurs étant amené à traiter l'information (techniciens d'information médicale, Assistants de recherche clinique, Techniciens de recherche clinique) sous la responsabilité d'un médecin de l'établissement.

Pour la réalisation d'études de faisabilité ou la production de tableaux de bord, l'exploitation des données est opérée par l'unité hospitalo-universitaire responsable de l'exploitation de l'entrepôt (Unité fouille de données - UFD) qui transmet au demandeur des décomptes et des agrégats floutés en cas de petits effectifs.

Pour une étude nécessitant le recueil de données individuelles, par exemple constitution d'une cohorte ou pré-recrutement dans un essai clinique, l'unité effectue une première requête sur l'entrepôt global à partir des critères d'éligibilité (par exemple une pathologie, une tranche d'âge, un ensemble d'unités fonctionnelles). Un sous entrepôt (Datamart) est constitué avec les documents résultats de la requête. Il est mis à disposition de l'investigateur de l'étude afin qu'il

repère sa cohorte et puisse recueillir et exploiter les données. A ce stade et parce que l'étude ne nécessite pas de connaître l'identité des patients, les données demeurent dé-identifiées.

Pour une étude où l'investigateur a besoin de recontacter les patients (par exemple pour proposer leur participation à un essai clinique), l'identité des patients sélectionnés est rétablie par l'UFD. Les patients qui se sont opposés à être recontactés sont alors exclus.

4.2 L'information des patients et leur accord à la réutilisation de leurs données de santé

Aux Etats unis, dans les centres hospitaliers où l'entrepôt I2b2 a été implanté, une note d'information est remise au patient [40]. Elle stipule que les données de son dossier médical seront exploitées pour la recherche sous forme dé-identifiée.

Il est à préciser que le patient signe un simple accusé de réception mais pas une autorisation. En effet, dans le cas où le traitement concerne des données dé-identifiées, l'accord du patient n'est pas requis pour la réutilisation de ses données de santé.

A Stanford, l'entrepôt STRIDE a été autorisé par l'IRB avec une dérogation à l'obtention du consentement du patient pour l'intégration de ses données de soins dans l'entrepôt [2].

Sur les deux sites, l'accès à des données identifiantes nécessite soit l'accord du patient, soit une dérogation au consentement du patient obtenu auprès d'un IRB.

Au CHU de Rennes il est prévu de joindre une note d'information avec le livret d'accueil selon le modèle du guide de bonnes pratiques pour assurer la confidentialité des données de soins lors de leur réutilisation pour des traitements biostatistiques [30].

Au Canada, un rapport a été commandité en 2009 par les autorités sur l'utilisation des données du dossier patient pour la recherche en santé [48]. Des propositions d'évolution du cadre réglementaire relatif à l'information du patient sont avancées. L'information serait collective et diffusée par les établissements ou les diverses structures de santé fréquentées par les patients, sous forme de brochures, affiches ou vidéo.

Le type d'accord du patient pourrait être conditionné par le domaine d'utilisation des données.

Par exemple, l'accord du patient est non requis pour des besoins de santé publique concernant la surveillance sanitaire, la pharmacovigilance et la matériovigilance.

Le recueil de l'opposition éventuelle du patient est requis pour les études relatives à la qualité des soins. Enfin le recueil de l'accord du patient à être recontacté est nécessaire pour les essais cliniques ou les études épidémiologiques comportant des inclusions.

Pour les biobanques, l'accord formalisé du patient est nécessaire dès lors qu'un prélèvement est réalisé et que les données le concernant sont recueillies.

Le recueil du choix du patient pourrait se faire au travers d'un portail web, dans l'esprit du Dossier Médical Personnel. Il serait également utilisé pour le retour d'informations au patient (indication des projets de recherche où ses données sont utilisées, résultats des études).

S'il est nécessaire de contacter le patient, cela se passe en deux temps. Dans un premier temps le patient est contacté par une personne habilitée à accéder à son dossier afin de savoir s'il est d'accord pour participer au projet de recherche. Dans l'affirmative, un membre de l'équipe de recherche le contacte pour recueillir son consentement éclairé.

5 Discussion

5.1 Une procédure d'information et de recueil de l'opposition difficile à mettre en oeuvre

La réutilisation des données de santé ne peut se faire sans en avoir informé le patient.

Or, la finalité précise d'un traitement, les données en faisant l'objet, les destinataires ne sont pas connus au moment du recueil initial. L'information individuelle des patients est difficile à mettre en oeuvre a posteriori.

Cette information ne peut être que générale et collective. Elle devrait figurer dans le livret d'accueil remis à tout patient hospitalisé. Au delà, à l'instar des propositions canadiennes, une campagne publique d'information, par exemple par voie de presse, pourrait être mise en place.

Sur les plans éthique et réglementaire, pour les études non interventionnelles, il est également nécessaire de recueillir auprès du patient son éventuelle opposition à réutiliser ses données ou à être recontacté [49]. L'exploitation des EDBM rentre dans ce cadre.

Pour en tenir compte au moment du traitement des données, cette opposition devrait être enregistrée dans le Système d'information de l'établissement.

En cas de recontact du patient, il serait préférable qu'il se fasse par un praticien intervenant dans sa prise en charge.

Le Dossier médical Personnel (DMP) pourrait être le lieu d'information et d'enregistrement de l'opposition du patient. L'opposition serait ensuite être transmise aux Système d'Information des établissements de santé. Les dernières orientations limitant le périmètre du DMP aux affections chroniques rendent inopérante cette solution. Par ailleurs, le recueil de l'opposition peut être source de biais dans l'exploitation des données [50]. La représentativité de la population n'est pas assurée. En effet, les patients qui s'opposent à la réutilisation de leurs données peuvent avoir un profil particulier.

5.2 Problématiques liée à la dé-identification

5.2.1 Des méthodes de dé-identification à améliorer

La réutilisation des données de santé serait possible sans formalité particulière, dans la mesure où elles sont anonymes. Or, une anonymisation automatique stricte est quasiment impossible à garantir. Cependant, les techniques de dé-identification permettent de s'en approcher. Elles devraient être systématiquement mises en œuvre pour permettre la réutilisation des données patient pour la recherche. Le traitement simple qui consiste à occulter les traits d'identité à partir des informations administratives du SIH n'est pas suffisant. En effet, Grouin [32] dans son travail mené sur l'anonymisation de compte-rendus de cardiologie retrouve 25% de documents avec persistance de données personnelles en utilisant cette technique. Les méthodes de dé-identification des données textuelles apportent un gain supplémentaire. Les algorithmes existants ont démontré de bonnes performances [34, 35]. Ils restent toutefois à améliorer, voire à combiner, pour la langue française [32]. Des travaux récents effectués à partir de MEDINA tendent à privilégier les méthodes d'apprentissage [51].

Il existe aussi un risque dégradation de l'exploitabilité de l'EDBM en raison de la suppression

possible de données d'intérêt en cas de surfiltrage, d'une moins bonne lisibilité et compréhensibilité des textes du fait du cryptage ou encore en l'absence de conservation de la temporalité des évènements [33].

5.2.2 Un besoin de certification des outils de dé-identification

A partir de quand peut-on considérer que la probabilité de remonter à l'identité du patient est suffisamment faible ? Cette question n'a pas reçu de réponse univoque. Il est proposé de demander au chercheur un engagement à ne pas essayer de ré-identifier le patient. A l'heure actuelle, concernant la dé-identification de données textuelles, aucune méthode n'a été validée ni certifiée à un échelon national. Par exemple, aux Etats Unis, l'algorithme de dé-identification dans l'entrepôt I2B2 est validé par l'IRB local. En France, dans la mesure où la dé-identification des dossiers médicaux serait la condition à leur utilisation pour la recherche, des directives plus précises sur les outils à utiliser devraient être émises.

5.2.3 Le risque de réidentification

En cas d'études longitudinales où sont analysées les trajectoires de soins du patient, le risque de réidentification est accru [33]. La combinaison d'informations non identifiantes prises isolément peut conduire à une description unique reliée à un patient si on a connaissance par ailleurs de certaines informations le concernant. Cette question a été évoquée dans le rapport de J. L Bras sur la gouvernance et l'utilisation des données de santé, lorsque sont examinées les conditions d'ouverture de la base SNIIRAM [52]. Cette problématique existe dans les EDBM puisque ceux-ci contiennent tous les épisodes de soins hospitaliers. Elle est d'autant plus prégnante pour les EDBM mettant en relation les données cliniques et les données génomiques. Certaines séquences sont d'une part liées à des maladies et sont d'autre part considérées comme identifiantes car caractérisant un seul patient, des informations médico-administratives pourraient permettre de remonter à l'identité du patient. Dans un article publié par *Science* en 2013 [53] les auteurs ont démontré qu'il était possible de ré-identifier un patient à partir d'informations contenues dans des bases publiques, une base de séquençage du génome et une base de généalogie. Avec l'avènement du big data et le croisement de données multi-sources peut-on encore aujourd'hui garantir l'anonymat lors des traitements des données ?

5.2.4 Un chaînage des données patient souvent nécessaire

Pour permettre des croisements avec d'autres sources de données externes (Base SNIIRAM par exemple) un numéro unique anonyme ou pseudonyme doit être généré. Des solutions ont déjà été proposées pour les études épidémiologiques multicentriques ou le recoupement de plusieurs études [54, 55]. Elles mettent en œuvre des fonctions de hachage et impliquent un tiers de confiance. La question de l'utilisation du NIR[6] et de son encadrement par un tiers de confiance est actuellement débattue au sein de la commission sur l'ouverture de la base SNIIRAM. La CNIL y serait favorable et pourrait par ailleurs émettre des recommandations de bonnes pratiques sur les techniques d'anonymisation des données [56]. Contrairement à l'Allemagne

[6] Toute personne née en France métropolitaine et dans les départements d'outre-mer (DOM) est inscrite au répertoire national d'identification des personnes physiques (RNIPP). L'inscription à ce répertoire entraîne l'attribution du numéro d'inscription au répertoire (NIR)

[37] ou le Royaume Unis [23-44], la France ne dispose pas à ce jour de service de pseudo-anonymisation accessible pour les chercheurs. Une procédure simple pour la création du numéro anonyme serait utile. Ceci permettrait de croiser les données avec d'autres sources comme des bases de données de recherche clinique à des fins de médecine translationnelle [57] ou de pharmacovigilance ou encore avec des bases de données épidémiologiques ou les registres.

5.2.5 Répondre au besoin de réidentification

Lorsqu'il est nécessaire de recontacter le patient, par exemple pour une étude prospective ou une inclusion dans un essai clinique, un lien doit être conservé avec les données d'identité. Dans un entrepôt, il est possible également de crypter les données identifiantes. Les clefs sont conservées à part. La gestion des liens permettant de remonter à l'identité ou des clefs de cryptage pourra être assurée par un tiers de confiance interne à l'établissement.

5.3 Aspects éthiques et réglementaires

Plusieurs questions se posent. L'accès à des données à caractère personnel nécessite-t-il l'accord du patient ? Le patient peut-il s'opposer à la réutilisation de ses données recueillies initialement pour soins ? L'intérêt individuel primerait-il sur l'intérêt collectif ?

K.W Goodman [58] estime que si les médecins sont investis d'une mission de santé publique (ex : déclaration de maladies obligatoires, causes de décès), les patients doivent aussi y contribuer.

Il semble cependant licite, qu'en vertu du respect de la vie privée [29] le patient ait un droit de regard sur l'accès à ses données personnelles.

Par contre la réutilisation de données ne permettant pas d'identifier le patient ne devrait pas être soumise à condition. On rappelle que pour la base SNIIRAM, s'agissant de données non identifiantes, l'accord du patient n'est pas requis pour son exploitation.

La directive européenne 95/46/CE est en cours de révision, et devrait être remplacée par un règlement sur la protection des données [59]. La mise en œuvre d'une pseudonymisation lors de traitement à caractère personnel notamment dans le cadre de la recherche médicale serait la règle.

Il faudra mieux définir la place de l'exploitation des EDBM dans ces dispositifs juridiques.

Concernant les démarches auprès de la CNIL, la mise en œuvre d'un EDBM au sein d'un établissement de santé nécessite d'effectuer une déclaration normale.

Ces aspects sont actuellement abordés dans la communauté d'informatique médicale à l'échelle internationale. Au niveau européen, trois principes sont retenus : la nécessité de définir une réglementation et un encadrement et d'informer pleinement les patients sur la réutilisation de leurs données de santé [61]. L'implication des citoyens et des patients est considérée comme obligatoire. La question d'un tiers de confiance pour le contrôle des accès aux données et la mise à disposition de celles-ci est posée.

5.3.1 Aspects déontologiques

Sur le plan déontologique, quel est le droit de regard des praticiens ayant produit les informations sur leur réutilisation ?

Le dossier médical est en fait une copropriété entre l'établissement, le patient et le médecin qui a produit les informations. Le médecin et l'établissement en sont dépositaires, le patient dispose

d'un droit d'accès [60].

Il est de bonne pratique d'informer et de s'assurer de l'accord des praticiens avant réutilisation des informations dont ils sont à l'origine.

Une évolution à l'échelle européenne du rôle du patient dans la gestion de ses données de santé pourrait avoir lieu, il en deviendrait le propriétaire et en contrôlerait ainsi l'accès [61].

5.3.2 Positionnement des patients

Les études menées sur ce sujet tendent à montrer que les patients sont réticents à la communication des données les concernant, même dé-identifiées.

Cette réserve concerne particulièrement l'accès ou la transmission de leurs données aux compagnies d'assurances ou à l'industrie pharmaceutique et de manière générale, chaque fois qu'il y a un intérêt commercial en jeu [62, 63].

Une majorité d'entre eux sont cependant favorables à la mise à disposition de leurs données aux équipes de recherche académiques et considèrent que le bénéfice l'emporte sur le risque d'atteinte à la vie privée. Leur principale crainte concerne les failles de sécurité des systèmes conduisant à des accès illicites [62, 63].

Dans une étude récente le choix des patients s'oriente vers le recueil de l'opposition et non du consentement à la réutilisation de leurs données de santé [63].

En France, le Collectif Inter associatif Sur la Santé (CISS) s'est prononcé pour le libre accès aux données publiques de santé anonymes constituées essentiellement par la base SNIIRAM [64].

D'un autre côté, les sites internet à destination des patients se développent comme le site américain, PatientLikeMe, les sites français Doctissimo ou Carenity. Ces sites emportent l'adhésion des patients et les positionnent de leur avis même comme partenaire dans l'évolution des connaissances médicales [66]

A leur initiative, les patients échangent des informations sur leur maladie et leur prise en charge. Les données ainsi collectées sont ensuite analysées (par exemple pour la recherche d'effets secondaires [67]) et peuvent être réutilisées pour la recherche clinique à titre payant. Les patients peuvent alors être sollicités, s'ils y consentent, pour répondre à des questionnaires ou participer à des études.

5.4 Sécurité des données dans les EDBM

A notre sens, la dé-identification des données dans les EDBM apporte de la confiance pour les patients et pour les utilisateurs lors de l'exploitation au sein des établissements en garantissant au mieux la confidentialité des donnés de santé.

Un risque de réidentification subsiste cependant. Certaines fonctionnalités renforcent ce risque, notamment les interrogations sur une courte période, la visualisation du parcours de soins du patient, les interrogations multi-critères sur des sous entrepôts.

On peut restreindre l'accès à des personnes que l'on considère de confiance, formées à la sécurité des informations de santé. Afin d'exploiter ces nouveaux outils dans toutes leurs fonctionnalités, un compromis devra être trouvé entre confidentialité et intérêt pour la recherche médicale.

L'avis du patient devra être pris en compte dès que le risque de réidentification est trop important.

La traçabilité des accès est assurée par étude. Elle concerne les utilisateurs et les requêtes effectuées.

Elle est renforcée lorsque les interrogations sont réalisées par les utilisateurs eux-mêmes, la multiplication des requêtes et des critères augmentant le risque de réidentification.

Sur le plan déontologique, la transparence est la règle, les études réalisées à partir de l'EDBM sont portées à la connaissance des médecins de l'établissement.

5.5 Propositions d'une organisation et d'une procédure pour l'exploitation des EDBM

Suite à l'analyse de la littérature des propositions peuvent être émises pour l'exploitation des EDBM dans le respect de la réglementation française et adaptés à l'organisation des établissements.

5.5.1 Les structures parties prenantes existantes

La responsabilité du traitement revient à la Direction de l'établissement de santé qui mettra en place (via la Direction des Systèmes d'Information) les mesures pour assurer la sécurité du système et effectuera la déclaration de l'EDBM pour une exploitation interne à l'établissement.

La Direction de la recherche porte à la connaissance des médecins et des soignants de l'établissement les modalités d'accès à l'EDBM.

La Direction des relations avec les usagers est l'interlocuteur des patients en cas d'opposition de leur part à la réutilisation de leurs données de santé.

La cellule communication est impliquée dans l'information du patient.

Le Directoire inscrit la mise en œuvre de l'EDBM dans le projet d'établissement et valide la politique d'exploitation de l'EDBM.

La Commission Médicale d'Etablissement informe les médecins et les soignants de l'établissement de la mise en œuvre de l'EDBM.

Le Département d'Information médicale s'assure du respect de la confidentialité des traitements lors de l'exploitation de l'EDBM.

Le Comité d'éthique contribue à définir les modalités d'exploitation de l'entrepôt respectant la protection des patients. En particulier ces questions deviendront cruciales avec l'intégration et le traitement de données OMICS, ou de biomarqueurs individuels.

La Commission des relations avec les usagers et la qualité de la prise en charge (CRUQPC) pourrait être un relais avec les associations de patients afin de valider les procédures d'information et de recontact. La CRUQPC est d'ailleurs amenée à évoluer tant sur sa composition avec une plus forte représentation des usagers, que sur ses missions notamment sur la garantie des droits des usagers [65].

5.5.2 Des structures dédiées à mettre en place

Un Comité de régulation de l'Exploitation des Données Patient (CREDoP) équivalent IRB

Il comprendra des représentants du corps médical, du DIM, du comité d'éthique, des patients, des directions de la recherche, de la Direction Informatique.

Ses missions consistent à :

valider le niveau d'accès à l'entrepôt et le traitement des données par étude

s'assurer que les règles déontologiques sont bien suivies,

vérifier que les autorisations réglementaires nécessaires ont été obtenues,

informer la communauté médicale des projets en cours dans un but de transparence.

Une Structure opérationnelle : le Centre de Données Cliniques (CDC)

Cette structure rassemble des acteurs compétents dans les domaines du traitement de l'information, en particulier médecins spécialistes en informatique médicale, data managers, techniciens de l'information, informaticiens, statisticiens.

La mission des CDC se déclinent ainsi:

Il développe et/ou maintient l'entrepôt de données, les outils d'exploitation et les bases de données patient en lien avec la recherche

Il instruit les demandes pour le CREDoP

Il réalise les extractions et le traitement des données et en fournit les résultats

Il fournit l'accès aux données patient individuelles, dé-identifiées ou non en fonction des demandes.

Il produit les indicateurs relatifs aux traitements et aux accès (reporting)

Il contribue avec le CREDoP à la rédaction de la politique d'exploitation de l'EDBM

5.5.3 Documents cadres

Politique d'exploitation et droit d'accès

Elle indiquera les modalités d'organisation des accès et d'affectation des droits d'accès.

Il faudra définir par qui sont réalisées les interrogations de l'EDBM, personnel d'une structure dédiée (par exemple de type CDC) ou médecins praticiens eux-mêmes.

Les cas d'usage sont définis : étude de faisabilité, pre-screening, constitution de cohortes. Ils déterminent le type de données visibles pour l'utilisateur, agrégées dans le premier cas, individuelles dans les autres cas, structurées ou textuelles selon le type de requête, nominatives ou dé-identifiées. La demande d'accès à des données nominatives doit être justifiée.

Des principes sont édictés concernant le périmètre des données interrogeables et consultables.

Les accès sont établis pour les professionnels de santé senior de l'établissement, les autres utilisateurs (interne, ARC, TEC) interviennent sous la responsabilité d'un professionnel de santé senior habilité à accéder au dossier du patient de par sa relation de soins avec les patients ou de par une mission confiée par l'établissement. Ce dernier est le plus souvent le responsable de l'étude ou l'investigateur principal.

Les accès à l'entrepôt se font sous son contrôle dans son service ou au sein de la structure chargée de l'exploitation de l'entrepôt.

Les utilisateurs n'accèdent par défaut qu'à des données dé-identifiées.

Pour les interrogations multi-services ou transversales sur l'établissement, la transparence est la règle avec une information des responsables médicaux concernés ou de la communauté médicale ou soignante de l'établissement.

Formulaire de demande d'accès

Il comprend les renseignements suivants : Nom du responsable de l'étude, type d'usage (faisabilité, pre-screening, repérage population, extraction données), type d'étude et objectif, justification pour la demande d'informations nominatives (recontact du patient, accès au dossier patient ….), accord des responsables médicaux pour les interrogations multi-services, autorisations des autorités compétentes jointes (études interventionnelles ou multicentiques), critères d'interrogation, professionnels de l'établissement utilisateurs de l'entrepôt pour cette étude (nom et fonction), période de l'étude, type de données, dans le cas d'extraction de données les informations souhaitées.

Charte d'utilisation

Tout utilisateur de l'EDBM s'engage à la respecter. Par là même il s'engage à respecter la confidentialité des informations auxquelles il a accès, à protéger les fichiers qu'il constitue et à les déclarer auprès du Correspondant Informatique et Libertés de l'établissement en cas de traitement de données à caractère personnel. Il s'engage également à ne pas essayer de réidentifier le patient.

Note d'information patient

La note d'information au patient explicite les besoins de réutilisation des données du dossier médical pour la recherche. Elle précise que le patient peut s'y opposer et que cela en l'occurrence ne modifie pas la qualité de sa prise en charge. Enfin, que les analyses sont réalisées de façon confidentielle dans le respect du secret médical.

5.5.4 Procédures

Procédure auprès des patients

La note d'information du patient est insérée dans le livret d'accueil.

L'opposition éventuelle du patient est enregistrée dans le système d'information pour être prise en compte dans l'exploitation de l'entrepôt. Le retour aux données identifiantes n'est alors pas possible.

Procédure auprès des utilisateurs

Il est souhaitable que les professionnels, utilisateurs de l'entrepôt soient sensibilisés à la sécurité des informations de santé et formés à l'utilisation des outils d'exploitation. Ces aspects devraient faire l'objet d'une formation initiale universitaire (par exemple dans le cadre de la certification C2I - 2 métiers de la santé) ou continue.

Circuit d'exploitation

Les demandes d'exploitation de l'entrepôt sont adressées au CDC qui joue le rôle de guichet unique. Elles y sont instruites sur le plan de leur faisabilité puis sont soumises à l'approbation du CREDoP. A partir des critères de sélection des patients, une requête est réalisée par le CDC. Un

sous entrepôt est constitué. Il est mis à disposition de l'équipe demandeuse sur une période déterminée. Le CDC peut apporter son appui méthodologique et statistique sur l'exploitation du sous entrepôt.

Avis CREDoP

L'avis du CREDOP concerne essentiellement les études transversales multiservices, ou multicentriques ou issues d'une demande extérieure à l'établissement.

L'avis est donné sur une période d'exploitation précise. L'accès est délivré sous la responsabilité d'un professionnel de santé interne à l'établissement.

Le CREDOP peut ne pas valider une demande si tous les prérequis ne sont pas réunis. Il doit informer le demandeur sur la raison du refus.

Procédure de reporting

La traçabilité des demandes ainsi que des interrogations et des consultations des sous-entrepôts doit être assurée et fait l'objet de rapports d'activité électroniques à destination du CREDOP, des services qui participent à l'étude, et des utilisateurs.

Chaque service possède un tableau de bord recensant l'ensemble des études (internes et externes au service), les utilisateurs et les données patients consultées du service.

5.5.5 Protection des données au sein de l'entrepôt

La consultation de l'ensemble des données de l'EDBM n'est pas possible, l'utilisateur ne peut voir que les informations d'un sous-ensemble de l'entrepôt résultat d'une requête. Une partition de l'entrepôt par service peut être prévue. Elle permet aux professionnels de santé appartenant à ce service d'exploiter les données des patients dont ils ont eu la charge.

La désidentification concerne à minima les traits d'identité des patients. Elle sera étendue de préférence à l'ensemble des noms et prénoms (professionnels de santé, famille), aux noms des établissements et services, dates, adresses (postales, électroniques), numéros (patient, séjour, établissement, service, téléphone). Les dates de naissance seront remplacées par l'âge, les autres dates par des dates fictives avec conservation de la temporalité. Les autres valeurs pourront être remplacées par des pseudonymes.

En fonction des cas d'usage, le retour aux données d'identification doit être possible. Un numéro patient sous forme de pseudonyme peut être créé pour permettre ensuite des retours aux données sources, par exemple pour des compléments d'informations sur un cas.

D'autre part des restrictions sont apportées lors des interrogations et la présentation des résultats. Il n'est pas possible d'interroger sur les traits d'identité d'un patient ou d'un professionnel (sauf à la demande de ce dernier). Hors besoins spécifiques (exemple : étude un jour donné) les interrogations sur une date précise ou sur un code postal ou code commune ne sont pas autorisées. L'affichage des statistiques agrégées respectent la borne de 10 patients, en deçà l'effectif est flouté et l'accès aux données individuelles n'est pas possible hors justification (ex : maladies rares).

6 Conclusion

La réexploitation secondaire des données patient représente actuellement un enjeu majeur pour la recherche en santé. Les EDBM contribuent à répondre à ce besoin et sont en pleine évolution : intégration de données OMIC, fédération d'entrepôts de données dont l'exploitation multicentrique devient nationale voire internationale, intégration ou couplage avec des données issues de bases nationales (comme la SNIIRAM) ou des données non médicales (par exemple environnementales), intégration des données des essais cliniques.

Au total nous connaissons une évolution technologique sans précédent, qui permet d'exploiter des données de santé multidomaines et à caractères personnel de façon multi échelle. A notre sens, la mise en œuvre de ces technologies au sein des établissements de santé doit s'accompagner d'une démarche claire et pragmatique. Cette démarche doit permettre d'encadrer strictement l'usage de ces outils sans toutefois entraver la recherche. C'est le sens des propositions que nous formulons dans ce travail.

Nous avons vu que sur le plan national le cadre juridique et les règles de bonnes pratiques éthiques et déontologiques doivent évoluer pour s'adapter à ce nouveau paradigme.

Plusieurs rapports parlementaires soulignent le besoin d'évolution des règles de protection des données dans les grandes bases biomédicales. Là aussi, il sera nécessaire de décliner sur le plan opérationnel les grands principes qui auront été définis et de les confronter à la réalité du terrain.

Référencess

[1] LOI n° 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel et modifiant la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

[2] Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2009;2009:391-395.

[3] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 26 févr 2010;17(2):124-130.

[4] Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 2009;48(1):38-44.

[5] Soto-Rey I, Bache R, Dugas M, Fritz F. Query engine optimization for the EHR4CR protocol feasibility scenario. Stud Health Technol Inform. 2013;192:1080.

[6] Cuggia M, Bayat S, Garcelon N, Sanders L, Rouget F, Coursin A, et al. A full-text information retrieval system for an epidemiological registry. Stud Health Technol Inform. 2010;160(Pt 1):491-495.

[7] Osmont M-N, Cuggia M, Polard E, Riou C, Balusson F, Oger E. [Use of the PMSI for the detection of adverse drug reactions]. Thérapie. août 2013;68(4):285-295.

[8] Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform. 2013;192:572-575.

[9] Bahl V, McCreadie SR, Stevenson JG. Developing dashboards to measure and manage inpatient pharmacy costs. Am J Health-Syst Pharm AJHP Off J Am Soc Health-Syst Pharm. 1 sept 2007;64(17):1859-1866.

[10] Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J Am Med Inform Assoc JAMIA. avr 2012;19(2):181-185.

[11] Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: the HEGP case. Stud Health Technol Inform. 2010;160(Pt 1):193-197.

[12] Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU. Unlocking Data for Clinical Research – The German i2b2 Experience: Appl Clin Inform. 30 mars 2011; 2(1):116-127.

[13] Stanford Center for Clinical Informatics - Stanford School of medicine. Access to Clinical Data - Services https://clinicalinformatics.stanford.edu/services/clinicaldata.html (consulté le 16 mars 2014).

[14] Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584-588.

[15] US departement of health and human services National Institutes of health. HIPAA Privacy Rule and Its Impacts on Research http://privacyruleandresearch.nih.gov/pr_08.asp#8a (consulté le 16 mars 2014)

[16] legislation.gov.uk. Data Protection Act 1998 http://www.legislation.gov.uk/ukpga/1998/29/contents (consulté le 15 mars 2014)

[17] Claudot F, Alla F, Fresson J, Calvez T, Coudane H, Bonaiti-Pellie C. Ethics and observational studies in medical research: various rules in a common framework. Int J Epidemiol. 1 août 2009;38(4):1104-1108.

[18] Bankert EA. Institutional Review Board: Management and Function. Jones & Bartlett Learning; 2006. 568 p.

[19] Health and Social Care Information Center . Caldicott Guardians. http://systems.hscic.gov.uk/infogov/caldicott (consulté le 18 janvier 2014)

[20] Commission nationale informatique et libertés. Guide des professionnels de santé Edition 2011. http://www.cnil.fr/fileadmin/documents/Guides_pratiques/CNIL-Guide_professionnels_de_sante.pdf (consulté le 23 septembre 2013).

[21] Official Journal of the European Communities. Directive 95/46/EC of the european parliament and of the council. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:FR:PDF (consulté le 23 septembre 2013).

[22] Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J Investig Med Off Publ Am Fed Clin Res. janv 2010;58(1):11-18.

[23] Powell J, Buchan I. Electronic Health Records Should Support Clinical Research. J Med Internet Res. 14 mars 2005;7(1):e4.

[24] National Health Service Connecting for health. http://www.connectingforhealth.nhs.uk/systemsandservices/pseudo/pipwhitepaper.pdf (consulté le 18 janvier 2014).

[25] National Health Service Connecting for health. http://www.connectingforhealth.nhs.uk/systemsandservices/pseudo/ref3deident.pdf (consulté le 18 janvier 2014)

[26] Agence technique de l'information sur l'hospitalisation. Présentation Programme de médicalisation des systèmes d'information en médecine, chirurgie, obstétrique (PMSI MCO). http://www.atih.sante.fr/mco/presentation?secteur=MCO (consulté le 15 mars 2014)

[27] El Emam K, Jonker E, Arbuckle L, Malin B. A Systematic Review of Re-Identification Attacks on Health Data. Scherer RW, éditeur. PLoS ONE. 2 déc 2011;6(12):e28071.

[28] Commission nationale de l'informatique et des libertés. L'état des lieux en matière de procédés d'anonymisation. http://www.cnil.fr/documentation/fiches-pratiques/fiche/article/letat-des-lieux-en-matiere-de-procedes-danonymisation/ (consulté le 23 septembre 2013)

[29] Code de la santé publique - Article L1110-4.

[30] EMOIS 2014. http://emois.org/images/EMOIS2014/soumission_resp.pdf (consulté le 13 avril2014).

[31] Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol. 2010;10(1):70.

[32] Grouin C. Thèse de doctorat de l'Université Pierre et Marie Curie. Spécialité Informatique Biomédicale. Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique. http://tel.archives-ouvertes.fr/docs/00/84/86/72/PDF/these_grouin.pdf (consulté le 22 mars 2014)

[33] Tamersoy A, Loukides G, Nergiz ME, Saygin Y, Malin B. Anonymization of Longitudinal Electronic Medical Records. IEEE Trans Inf Technol Biomed. mai 2012;16(3):413-423.

[34] Uzuner O, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assoc. 28 juin 2007;14(5):550-563.

[35] Deleger L, Molnar K, Savova G, Xia F, Lingren T, Li Q, et al. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J Am Med Inform Assoc. 2 août 2012;20(1):84-94.

[36] Kushida CA, Nichols DA, Jadrnicek R, Miller R, Walsh JK, Griffin K. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med Care. juill 2012;50 Suppl:S82-101.

[37] Faldum A, Pommerening K. An optimal code for patient identifiers. Comput Methods Programs Biomed. juill 2005;79(1):81-88.

[38] Uzuner O, Sibanda TC, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med. janv 2008;42(1):13-35.

[39] Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. Int J Med Inf. déc 2010;79(12):849-859.

[40] Chazard E, Mouret C, Ficheur G, Schaffar A, Beuscart J-B, Beuscart R. Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records. Int J Med Inf. 7 déc 2013;

[41] Grouin C, Zweigenbaum P. Automatic de-identification of French clinical records: comparison of rule-based and machine-learning approaches. Stud Health Technol Inform. 2013;192:476-480.

[42] Gateshead Health NHS Foundation Trust. Pseudonymisation Policy. http://www.qegateshead.nhs.uk/sites/default/files/users/user1/IG08%20Pseudonymisation%20Policy.pdf (consulté le 23 septembre 2013).

[43] Health and Social Care Information Centre. Who we are and what we do. http://www.hscic.gov.uk/whoweare (consulté le 15 mars 2014).

[44] Health and Social Care Information Centre. Data Linkage and Extract Service. Disponible sur: http://www.hscic.gov.uk/dles (consulté le 15 mars 2014).

[45] Institut des données de santé. Etudes et recherche - Présentation du SNIIRAM. http://www.institut-des-donnees-de-sante.fr/upload/06_etudes_recherches/01_form/SNIIRAM_Presentation_20131014.pdf (consulté le 16 mars 2014).

[46] Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc. 7 oct 2011;18(Suppl 1):i103-i108.

[47] Riou C, Cuggia M, Garcelon N. Comment assurer la confidentialité dans les entrepôts de données biomédicaux ? Colloq Adelf-Émois Système Inf Hosp Épidémiologie. mars 2012;60, Supplement 1(0):S19-S20.

[48] Willison DJ. Use of Data from the Electronic Health Record for Health Research - current governance challenges and potentialapproaches. http://www.priv.gc.ca/information/research-recherche/2009/ehr_200903_e.pdf (consulté le 20 septembre 2013).

[49] Claudot F, Fresson J, Coudane H, Guillemin F, Demoré B, Alla F. Recherche en épidémiologie clinique : quelles règles appliquer ? Rev DÉpidémiologie Santé Publique. févr 2008;56(1):63-70.

[50] Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. BMJ. 2009;338:b866.

[51] Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform. 2013 Dec 29. pii:S1532-0464(13)00205-0.

[52] Ministère des Affaires sociales et de la Santé. Rapport sur la gouvernance et l'utilisation des données de santé. http://www.social-sante.gouv.fr/IMG/pdf/Rapport_donnees_de_sante_2013.pdf (consulté le 18 janvier 2014).

[53] Identifying personal genomes by surname inference. Gymrek M1, McGuire AL, Golan D, Halperin E, Erlich Y. Science. 2013 Jan 18;339(6117):321-4.

[54] Lo Iacono L. Multi-centric universal pseudonymisation for secondary use of the EHR. Stud Health Technol Inform. 2007;126:239-247.

[55] Quantin C, Fassa M, Coatrieux G, Riandey B, Trouessin G, Allaert FA. Linking anonymous databases for national and international multicenter epidemiological studies: A cryptographic algorithm. Rev DÉpidémiologie Santé Publique. févr 2009;57(1):e1-e6.

[56] Dépêche APM du 4 décembre 2014, éditeur. Open data : la CNIL souscrit à un grand nombre de propositions du rapport Bras. 2014.

[57] Szalma S, Koka V, Khasanova T, Perakslis ED. Effective knowledge management in translational medicine. J Transl Med. 2010;8:68.

[58] Goodman KW. Ethics, information technology, and public health: new challenges for the clinician-patient relationship. J Law Med Ethics J Am Soc Law Med Ethics. 2010;38(1):58-63.

[59] Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard yo the processong of personal data and on the free movement of such data (General Data Protection Regulation). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:PDF (consulté le 23 septembre 2013)

[60] Haute Autorité de Santé. Evaluation des pratiques professionnelles dans les établissements de santé Dossier du patient Réglementation et recommandations. Disponible sur: http://www.has-sante.fr/portail/upload/docs/application/pdf/2009-08/dossier_du_patient_-_fascicule_1_reglementation_et_recommandations_-_2003.pdf (consulté le 18 janvier 2014).

[61] Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: A transnational perspective. Int J Med Inf. janv 2013;82(1):1-9.

[62] Perera G, Holbrook A, Thabane L, Foster G, Willison DJ. Views on health information sharing and privacy from primary care practices using electronic medical records. Int J Med Inf. févr 2011;80(2):94-101.

[63] Hill EM, Turner EL, Martin RM, Donovan JL. « Let's get the best quality research we can »: public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. BMC Med Res Methodol. 2013;13:72.

[64] Dépêche APM du 28 janvier 2013, éditeur. Le CIss réclame un libre accès aux données de santé. 2013.

[65] Ministère des Affaires sociales et de la Santé. Rapport à la ministre des Affaires sociales et de la Santé : Pour l'an II de la Démocratie sanitaire. [cité 21 mars 2014]. Disponible sur: http://www.sante.gouv.fr/IMG/pdf/Rapport_DEF-version17-02-14.pdf (consulté le 21 mars 2014).

[66] deBronkart D. How the e-patient community helped save my life: an essay by Dave deBronkart. BMJ. 2013 Apr 2;346:f1990.

[67] Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu Symp Proc. 2011;2011:1019-26.

Adresse de correspondance

Christine Riou
Département d'Information Médicale CHU de Rennes
35033 Rennes Cedex, France
E-mail : christine.riou@chu-rennes.fr