Note d'étape sur le Health Data Hub, les entrepôts de données de santé et les questions éthiques posées par la collecte et le traitement de données desanté dites " massives "

Pierre Lombrail, Israël Nisand, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain, Bernard Baertschi, Anne Buisson, Catherine Cornu, François Hirsch, Christine Lemaitre, et al.

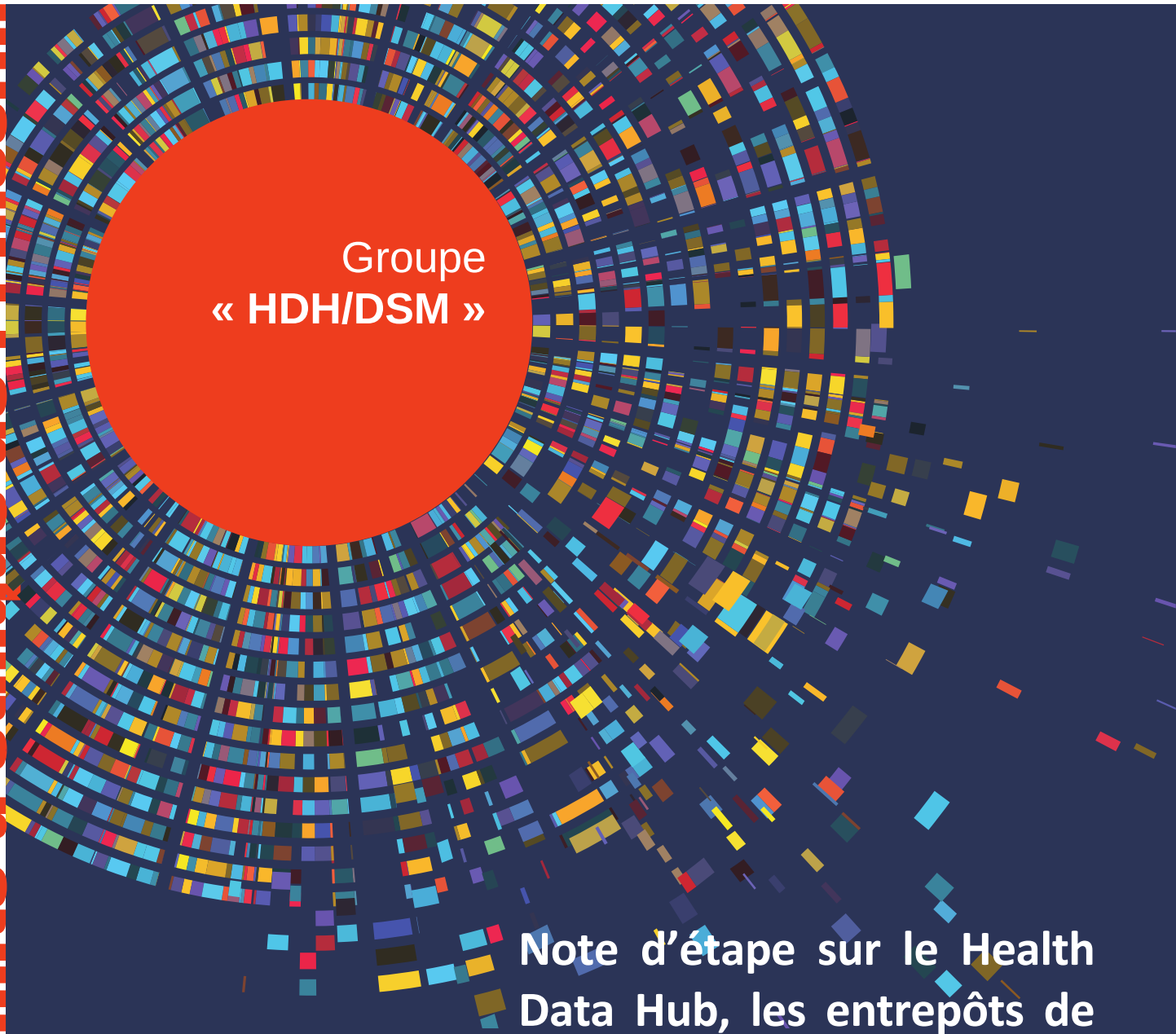
RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Inserm

La science pour la santé
From science to health

Comité d'éthique de l'Inserm



Groupe
« HDH/DSM »

Note d'étape sur le Health Data Hub, les entrepôts de données de santé et les questions éthiques posées par la collecte et le traitement de données de santé dites « massives »

Janvier 2022

Note d'étape sur le Health Data Hub, les entrepôts de données de santé et les questions éthiques posées par la collecte et le traitement de données de santé dites « massives ».

Pierre Lombrail, Israël Nisand, copilotes du « groupe de réflexion HDH/DSM » du CEI, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain et les membres du groupe : Bernard Baertschi, Anne Buisson, Catherine Cornu, François Hirsch, Christine Lemaitre, Sylvie Ledoux, Flavie Mathieu, Isabelle Rémy-Jouet, Yamina Sadani.

Le CEI a mis en place un groupe de travail en octobre 2020 du fait des interrogations soulevées par la décision de confier l'hébergement des données du Système national des Données de Santé (SNDS) rassemblées par le Health Data Hub (HDH ou PDS pour Plateforme des données de santé) à la société Microsoft à travers son « cloud » Azure. Le groupe a été amené rapidement à élargir sa réflexion à un ensemble plus vaste de questions éthiques soulevées par la collecte et le traitement de données dites « massives » pouvant s'apparenter de près ou de loin à des données de santé.

Depuis que le groupe a commencé son travail est paru le Décret n° 2021-848 du 29 juin 2021 relatif au traitement de données à caractère personnel dénommé « système national des données de santé » (SNDS) qui prévoit les modalités de gouvernance et de fonctionnement du système national des données de santé dont le périmètre a été étendu à de nouvelles catégories de données[i] par la loi n°2019-774 du 24 juillet 2019, telles que les données à caractère personnel des enquêtes dans le domaine de la santé, appariées avec des données du SNDS et régies par les dispositions de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques et il semble attendu que s'inscrivent au « catalogue » du HDH qui sera défini par arrêté, des bases de données constituées par des équipes INSERM (en sus du registre des causes médicales de décès d'ores et déjà intégré dans le périmètre du SNDS « historique »). Par ailleurs, outre les accès

ponctuels au SNDS, soumis à l'accomplissement préalable d'une formalité auprès de la CNIL, l'accès permanent au SNDS a été étendu à de nombreux organismes publics ou chargés d'une mission de service public, pour les besoins de leurs missions, dont font partie « les équipes de recherche de l'INSERM » (tout comme [CSP, art. R. 1461-13] « les équipes de recherche des centres hospitaliers universitaires et des centres de lutte contre le cancer », auxquelles leurs membres appartiennent souvent).

Il devient d'autant plus nécessaire de réfléchir à la nature et aux conditions de respect des obligations qui en découlent pour l'institution INSERM (politique générale et gouvernance formalisée d'un accès aux données, formation de l'ensemble des personnes concernées, support suffisant et compétent pour l'instruction des projets de recherche et la documentation interne de leur conformité réglementaire par des équipes aux moyens renforcés) et les équipes de recherche en termes de protection des droits des personnes qui participent à la recherche et de garantie de validité scientifique des travaux menés. Il y va de la pratique d'une recherche éthique et responsable mais plus globalement de démocratie en santé.

Dans cette note d'étape, nous faisons état de questionnements identifiés à la faveur d'une première série d'auditions et nous esquissons certaines pistes d'amélioration des politiques et procédures pour garantir le plus grand respect possible des droits des personnes participant à la recherche ainsi que la validité scientifique des travaux. A l'heure où certains font état d'une « urgence à faciliter l'accès » aux données de santé (« Les données de santé servent l'intérêt public, il y a urgence à en faciliter l'accès ». Tribune, Collectif, Le Monde 20 octobre 2021), il nous semble que loin de freiner la recherche, ces pistes sont de nature à garantir la confiance des participants à la recherche qui concourent au progrès de la science et par là-même, la pérennité d'une recherche de qualité. La réflexion du groupe va se poursuivre, pour mieux comprendre les usages dans différents contextes, avec une attention particulière à la mise en œuvre des techniques « d'intelligence artificielle ».

I. Des préoccupations initiales liées à l'hébergement du HDH sur le cloud Azure aux interrogations sur la protection des droits des personnes et l'intégrité scientifique liées à la constitution d'entrepôts de données de grande taille

La création du Health Data Hub (HDH/PDS) est l'aboutissement d'une longue histoire qui fait de la France le dépositaire d'un des plus volumineux entrepôts de données de santé au monde[ii]. Il s'agit d'un patrimoine unique en termes de capacité de production de connaissances et de potentiel d'optimisation du fonctionnement du système de soins[iii] dont l'existence même soulève des questionnements que le président du CEI formulait en ces termes dès 2016 dans un ouvrage consacré aux « Big Data »[iv] :

> *« L'utilisation des données massives en santé ou « Big Data » biomédicale illustre la tension éthique générée par d'une part, l'énorme potentiel de la méthode pour faire avancer la connaissance des origines, le diagnostic, le traitement et la prévention des maladies, et d'autre part, la sensibilité des informations relatives à la santé, couvertes en principe pour cela par le secret médical, et la vulnérabilité implicite générées par leur accessibilité, en partie liée au manque de formation du public au sens de ces données et à l'opacité des algorithmes utilisés pour les obtenir. »*

L'hébergement de la plateforme nationale des données de santé (« Health Data Hub ») par le cloud Azure de Microsoft a inquiété le CEI pour trois raisons au moins :

- opérateur privé (laissant craindre l'influence d'enjeux financiers sur les choix d'organisation et de valorisation de l'infrastructure potentiellement contraires à l'intégrité scientifique et à l'intérêt collectif ; crainte renforcée par les interrogations sur le modèle économique du HDH à terme s'il devient prisonnier de ces enjeux),
- de droit américain (craintes sur la protection des droits des personnes du fait du Cloud Act et de la réglementation FISA avec la possibilité de transfert des données hors du territoire pour transmission à des autorités requérantes nord-américaines),
- et choix d'une architecture centralisée vulnérable au « piratage » (alors que le dossier de préfiguration du HDH suggérait le choix d'une configuration de stockage répartie limitant les risques d'intrusion). Ce dernier point a été relevé depuis dans la délibération de la CNIL[v] qui fait suite à la publication du décret[vi] HDH[vii]. Il est d'autant plus saillant qu'elle est amenée à

relever que, « d'après les précisions apportées par le ministère, la PDS disposera d'une copie de la base principale, actuellement hébergée par la CNAM et que la base catalogue sera uniquement hébergée par la PDS ».

Le CEI est conscient des enjeux scientifiques, économiques et industriels, pharmaceutiques notamment, qui président au déploiement du HDH, mais il est vigilant à deux titres au moins :

- dans un contexte de concurrence exacerbée entre sociétés de services prétendant contribuer à l'optimisation du fonctionnement des systèmes de santé[viii] (voir le consortium Nvidia qui se met en place au Royaume-Uni[ix] ou la récente création de l'Alliance française des données en vie réelle dans notre pays[x]), la fragilité du modèle économique des plateformes qui se mettent en place inquiète quant à l'équilibre entre nécessaire retour sur investissement à terme (un des partenaires d'Agoria, « plateforme de collecte et d'analyse de données de santé au service d'une meilleure prise en charge thérapeutique des patients » impliquant notamment DocaPoste, parle de « plateforme marchande »[xi,xii]) et qualité des projets de recherche (et des services) ; au-delà, après la multiplication de scandales portant sur des accès / traitements frauduleux à des fins commerciales, le CEI s'inquiète du respect des droits des personnes (même s'il ne s'agit pas de consentement éclairé à proprement parler, l'effectivité du droit d'opposition à la réutilisation de données personnelles ne semble pas systématiquement garantie[xiii] ; il est préoccupé enfin par l'absence de reconnaissance du travail fourni pour la « production des données », tant par les bénévoles qui contribuent à leur collecte que par les chercheurs dont l'expertise permet de leur conférer du sens) ;
- un choc de paradigmes scientifiques émerge qui doit être soigneusement analysé : le modèle de production de connaissances dominant dans le monde biomédical est hypothético-déductif ; il part de l'hypothèse pour conduire un travail de réfutation, que ce soit au laboratoire ou dans la recherche épidémiologique ; le modèle des plateformes est inverse avec une collecte de données première, un croisement de sources multiples et la mise en évidence d'associations potentiellement signifiantes par la mise en œuvre de méthodes d'intelligence artificielle[xiv]. Ceci remet tout d'abord en question la définition même de ce qui peut être considéré comme une donnée de santé. Ceci entraine ensuite des débats épistémologiques ardus et une question éthique consiste à pouvoir garantir la validité des traitements de données et de l'interprétation des résultats produits. A supposer que ceux-ci soient toujours licites quant au respect des droits des personnes encore une fois. Les questions relatives à la mise en œuvre des techniques d'intelligence artificielle n'entrent pas dans le cadre de cette première note, elles seront étudiées dans un second temps et le sujet des algorithmes est identifié comme central[xv]. Sans aller jusqu'à dénoncer comme Eric Sadin un « antihumanisme radical[xvi] », nous sommes attentifs aux propos du président du comité d'éthique du CNRS,

Jean-Gabriel Ganascia, quand il déclare « il faut une réflexion forte sur les limites à imposer à l'IA[xvii] ». Relevons pour l'instant que la loi de bioéthique du 2 août 2021 (art. 17) vient encadrer le traitement algorithmique de données massives et la décision médicale en créant de nouvelles obligations en matière d'information et d'« explicabilité » pour pallier les dangers de perte de contrôle et de déresponsabilisation des utilisateurs professionnels (CSP, art. L.4001-3 nouveau).

Un premier échange au sein du comité fin 2020 a relevé certains des enjeux : « au-delà du HDH, c'est l'ensemble des collections de données de santé qui sont concernées, y compris celles issues des « entrepôts de données »[xviii] et des cohortes[xix] ; la qualité et l'intégrité des données sont centrales, tout comme le sont les questions de confidentialité (la pseudonymisation est-elle une protection suffisante) et de consentement (à quoi, pour combien de temps, sur la base de quelle information) ou de non-opposition des personnes concernées ; la centralisation se justifie-t-elle si elle permet la mise en place de procédures de sécurité aussi massives que les risques de forçage[xx] qu'elle fait courir quand il existe des méthodes d'analyse « distribuée » de données multi-sources[xxi] ; la complexification de l'accès au HDH handicape la recherche publique quand elle n'a pas les mêmes moyens d'investir que de grands opérateurs privés, or il faut s'inquiéter autant de traitements « par excès » que d'absence de traitements qui seraient source de progrès ; sans oublier l'enjeu éthique de la qualité des projets de recherche qui sont soumis au comité pour avis préalable à l'autorisation d'accès au HDH, le CESREES (Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé) ; comment articuler les comités scientifiques locaux, essentiels pour les producteurs de données et le rôle dévolu au CESRESS ; enfin, si pour l'une d'entre nous la vision de la vie bonne européenne n'est pas l'américaine, le RGPD (Règlement général sur la protection des données ; *voir encadré*) offre une base de réflexion /protection précieuse.

RGPD (règlement général sur la protection des données)

Le RGPD encadre, depuis le 25 mai 2018, le traitement des données personnelles sur le territoire de l'Union européenne et même au-delà dès lors que l'on cible des résidents européens[xxii]. Son champ d'application matériel exclut les données personnelles rendues anonymes de manière irréversible par un processus permettant de garantir que la personne concernée ne pourra pas être réidentifiée par la suite mais il inclut les données pseudonymisées (cf infra). Celles-ci restent en effet attachées à la personne concernée, par exemple à l'aide d'un identifiant, même si elles ne peuvent « plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires ». Le champ d'application géographique du RGPD s'étend aux organismes établis en Europe, que le traitement ait lieu ou non dans l'Union, mais pas seulement, dès lors que le traitement cible des résidents européens. En outre, la France a choisi d'adopter des spécificités locales dans son adaptation de la Loi Informatique et Libertés qui veulent que les règles nationales « s'appliquent dès lors que les personnes concernées résident en France, y compris lorsque le responsable du traitement n'est pas établi en France ».

Les principes fondamentaux de la protection des données nous intéressent particulièrement (les guillemets signalent les citations d'un opuscule de Aurélie Banck, déjà signalé) :

- L'objectif poursuivi par le responsable du traitement des données doit répondre à des finalités « déterminées » (ce qui exclut toute collecte de données au hasard ou à des fins préventives), « explicites » (c'est à dire communiquées à la personne concernée), « légitimes » par rapport à l'organisme mettant en œuvre le traitement ;

- Les données doivent être « traitées de manière licite, loyale et transparente au regard de la personne concernée ». La personne concernée doit consentir au traitement ou celui-ci doit être nécessité par des conditions précises spécifiées (l'exigence de loyauté et de transparence renvoie à l'information des personnes concernées et vise à éviter les traitements occultes ou cachés) ;

- Les données doivent être « adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées » (ce principe de minimisation des données, également appelé principe de proportionnalité, vise à garantir que l'ensemble des données collectées sont strictement nécessaires par rapport à l'objectif poursuivi et à exclure toute collecte réalisée au cas où ces données se révéleraient utiles a posteriori) ;

- Les données doivent être « exactes » et, si nécessaire, tenues à jour (ce qui est par exemple critique quand on traite les données du programme de médicalisation des systèmes d'information, PMSI hospitalier dont les nomenclatures et les algorithmes de classement des séjours changent régulièrement, ce qui nécessite des données de contexte pour interpréter correctement les résultats des traitements sur le temps long) ;

Les données doivent être « traitées de façon à garantir une sécurité appropriée des données à caractère personnel, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de mesures techniques ou organisationnelles appropriées » et adaptés aux risques (principes d'intégrité et de confidentialité). Particulièrement s'agissant de données « sensibles » ou portant sur des catégories dites « particulières» (« origine raciale ou ethnique, opinions … ainsi que le traitement de données génétiques ou biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne sont interdits ») sauf à pouvoir se prévaloir d'une des exceptions limitativement énumérées par les textes dont une concerne la recherche scientifique. Cette interdiction de principe n'interdit pas le traitement de ces données sensibles mais il faut pour pouvoir traiter ces données : pouvoir justifier d'une des exceptions légales à l'interdiction ; entourer le traitement de ces données sensibles de garanties appropriées (de fond et de procédure).

- Ces considérations sont développées, notamment pour ce qui regarde les caractéristiques du consentement des personnes concernées, qui doit être libre, spécifique (donné pour une finalité spécifique précise et de manière granulaire) et éclairé (recueilli notamment dans un langage clair, accessible et compréhensible). Parmi les données sensibles, des considérations particulières s'appliquent au NIR (numéro d'inscription au répertoire des personnes physiques) communément appelé numéro de sécurité sociale en raison de son caractère signifiant et des risques d'interconnexion qui s'y attachent.

- « Accountability » : le RGPD introduit un changement de paradigme majeur avec ce principe de responsabilité qui fait basculer d'un régime de formalité préalable statique (la déclaration / autorisation de traitement) à un régime de conformité globale dynamique. Ce principe se traduit par la définition de politiques de protection des données et de sécurité des systèmes d'information, d'un registre des activités de traitement et des violations de données et par la prise en compte des principes de *Privacy by design* et *Privacy by default*.

- On terminera ce rappel non exhaustif par une des nouveautés majeures introduites par le RGPD : l'analyse d'impact relative à la protection des données. Il s'agit d'une analyse des risques pour la vie privée des personnes concernées résultant de la mise en œuvre d'un traitement de données à caractère personnel quand ce risque peut être considéré comme élevé, ce qui est le cas des projets de recherche en santé humaine (sur la base de 9 critères spécifiés). Sa réalisation nécessite la collaboration étroite de tous les opérateurs concernés par le traitement [xxiii]. Elle sera systématiquement requise et doit être impérativement fournie à l'appui de la demande d'autorisation initiale en cas de :

• recherche médicale portant sur des patients et/ou sur des mineurs et incluant le traitement de leurs données génétiques ;

• constitution d'un registre ou d'une base de données (« entrepôt de données ou d'échantillons biologiques ») ayant vocation à être ouverts à la communauté de recherche.

I.1 Protection des droits des personnes relatifs au traitement de données à caractère personnel

La CNIL veille au respect de la protection des données et des droits des personnes en France selon des principes formulés dès la loi Informatique et Libertés de 1978, révisée en 2019 pour intégrer au droit français les évolutions introduites par le droit européen à travers le RGPD. Deux grands types d'exigences peuvent être distingués : intégrité et confidentialité des données d'une part, transparence des traitements d'autre part. Nous y ajoutons dans le cadre de la recherche le droit à la reconnaissance du travail réalisé pour permettre la constitution des bases de données, celui des contributeurs bénévoles qui prennent de leur temps pour répondre régulièrement à des questionnaires souvent longs, celui des chercheurs qui ont conçu les protocoles et procéderont aux analyses qui permettront de produire des connaissances nouvelles. Le tout dans un double mouvement, science ouverte et recherche participative.

I.1.1 Intégrité et confidentialité des données

Selon le RGPD : « Les données doivent être « traitées de façon à garantir une sécurité appropriée des données à caractère personnel, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de mesures techniques ou organisationnelles appropriées ». Nous ne ferons que rappeler les craintes soulevées par le choix d'hébergement effectué par le HDH. Elles nous semblent d'autant plus fondées qu'il existe des alternatives françaises d'hébergement, y compris une offre publique, le CASD (Centre d'accès sécurisé aux données). Il semble que le gouvernement revienne progressivement sur l'enjeu de « souveraineté » avec l'obligation pour les administrations de ne recourir qu'à des prestataires labellisés Cloud de confiance pour les services traitant les données de citoyens, d'entreprise ou d'agents publics. Ce domaine est extrêmement mouvant, tant sur un plan technique que juridique et une veille s'impose. Nous insisterons par ailleurs sur le fait qu'aucune procédure ne saurait garantir une absolue sécurité et confidentialité des données personnelles qu'il s'agisse de pseudonymisation voire même d'anonymisation (voir encadré *Données pseudonymisées et anonymisées*).

Données pseudonymisées et anonymisées

1. Qu'est-ce qu'une donnée pseudonymisée ?

Les données sont dites pseudonymes lorsque l'attribution à une personne concernée requiert le recours à des informations supplémentaires (Table de correspondance, clé de chiffrement, etc). Les données résultant d'une pseudonymisation sont donc considérées comme des données personnelles et leur traitement reste soumis aux principes de protection des données.

En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.). La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe.

Cette définition couvre différentes techniques couramment utilisées en matière de recherche en santé:

- le recours à une table de correspondance entre le jeu de données pseudonymes (codées) nécessaire aux analyses et les données d'identité conservées séparément, classiquement utilisée dans les essais cliniques ;

- chiffrement de données directement identifiantes pour les rendre incompréhensibles avec un secret qui permette de chaîner des données relatives à un individu et de suivre son parcours sans permettre de l'identifier.

La pseudonymisation est un traitement de données personnelles assurant la sécurité des données tout en préservant intégralement leur utilité. C'est une opération réversible, contrairement à l'anonymisation. En pratique, il est possible de retrouver l'identité des personnes dont les données ont été pseudonymisées.

Les textes encouragent l'utilisation de la pseudonymisation dans le cadre de la recherche scientifique (RGPD, art. 89). La pseudonymisation réduit en effet le risque de mise en corrélation d'un ensemble de données avec l'identité originale d'une personne concernée et concourt à ce titre, à la minimisation des risques pour les personnes.

Quelle que soit la technique de pseudonymisation appliquée, les informations permettant de mettre en relation les pseudonymes générés et les données directement identifiantes revêtent une sensibilité importante. Il convient de s'assurer que la confidentialité de ces éléments est assurée par des mesures techniques et organisationnelles appropriées. Ces informations ne doivent ainsi pouvoir être accédées que par des personnes autorisées et dans des conditions préalablement spécifiées.

2. Les données anonymes ou anonymisées ?

Les données anonymes ou anonymisées ne sont pas soumises aux législations de protection des données qu'elles soient anonymes initialement ou après une anonymisation par un traitement permettant de garantir que la personne concernée ne pourra pas être réidentifiée par la suite (données anonymisées). Travailler sur des données anonymes ou anonymisées permet donc de s'affranchir de la réglementation car la diffusion ou la réutilisation des données anonymisées n'a pas d'impact sur la vie privée des personnes concernées.

L'impossibilité d'identifier les personnes requiert une évaluation des risques au cas par cas. Elle s'apprécie au regard des moyens raisonnablement susceptibles d'être utilisés par le responsable de traitement ou par toute autre personne. Elle passe en pratique par la prise en considération du coût de l'identification, du temps nécessaire à celle-ci, des technologies disponibles, présentes mais aussi à venir. En cas de doute sur le caractère identifiant des données, il est recommandé de considérer les données comme identifiantes, jusqu'à la preuve du contraire.

Comment évaluer un processus d'anonymisation ?

Dans leur avis de 2014, les autorités de protection des données européennes définissent trois critères qui permettent de s'assurer qu'un jeu de données est véritablement anonyme :

- Non-individualisation : il ne doit pas être possible d'isoler un individu dans le jeu de données
- Non-corrélation : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
- Non-inférence : il ne doit pas être possible de déduire de façon quasi certaine de nouvelles informations sur un individu.

À défaut de remplir parfaitement ces trois critères, il doit être démontré, via une évaluation approfondie des risques d'identification, que le risque de ré-identification avec des moyens raisonnables est nul.

Les techniques d'anonymisation et de réidentification étant amenées à évoluer régulièrement, il est indispensable, d'effectuer une veille régulière afin de préserver, dans le temps, le caractère anonyme des données produites.

En pratique, l'anonymisation suppose donc :

- Une évaluation, au cas par cas, en tenant compte à la fois du contexte et du risque, de la technique ou de la combinaison de techniques d'anonymisation, étant entendu qu'aucune technique n'est infaillible, comme l'indiquait le G29 et ce que confirment les travaux de recherche en la matière,
- Une réévaluation régulière au regard de l'évolution des techniques(v),
- La destruction irréversible des données initiales(vi).

Selon le RGPD toujours, les données doivent être « traitées de manière licite, loyale et transparente au regard de la personne concernée »). Cette exigence est complexe à satisfaire pour plusieurs raisons :

- la première tient à la définition-même de ce qui peut être considéré comme une donnée de santé et qui conditionne pour une part le niveau d'information du citoyen quant à certains usages ;
- la seconde relève de la qualité des dispositifs d'information des personnes qui fonde leur capacité soit à donner leur consentement à l'utilisation et à la réutilisation de leurs données, soit d'exercer leur droit d'opposition voire d'effacement de certaines données personnelles.

La définition de données à caractère personnel (*voir encadré Les données de santé : une définition large*) semble juridiquement claire[xxiv], tout comme celle de traitement[xxv], mais celle de données de santé à caractère personnel se prête pourtant à des interprétations variables (le collectif SantéNathon est en désaccord avec le Conseil d'Etat par exemple sur le statut des données de rendez-vous médicaux enregistrées par le site Doctolib ; le seul qualificatif « médical » appliqué à ces données ne laisse pourtant guère de doute selon nous). Surtout, cette note ne porte que sur les données explicitement recueillies à des fins de recherche ou de soins / administration de la santé, et il nous restera à prendre en compte celles qui sont recueillies et transmises par toutes sortes d'objets « connectés » dont les personnes s'équipent dans une préoccupation de bien-être[xxvi] (« internet des objets ») ou à des fins explicites de suivi médical (avec un besoin de clarification des conditions de respect de la confidentialité des données quand des prestataires privés fournissent des conseils individualisés de suivi médical mobilisant des données collectées lors de soins par exemple)[xxvii].

Les données de santé : une définition très large

Les « données concernant la santé » sont définies comme « les données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne »).

L'état de santé s'entend de l'**état de santé physique, mentale, spirituelle et sociale présent, passé ou futur de la personne** et les données comprennent très largement :

- toute information sur l'identification du patient dans le système de soins,
- toutes les prestations de services de santé,
- des informations obtenues lors du test ou de l'examen d'une partie du corps ou d'une substance corporelle, y compris à partir de données génétiques et d'échantillons biologiques (données de santé « par destination ») ;
- toute information concernant, par exemple, une maladie, un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l'état physiologique ou biomédical de la personne concernée, indépendamment de sa source (considérant 35).

Sont concernées :

- les données qui permettent d'indiquer la pathologie dont peut être atteint un individu (données de santé « **par nature** ») ; antécédents médicaux, maladies, prestations de soins réalisés, résultats d'examens, traitements, handicap, etc.
- le croisement de données qui permettent de tirer une conclusion sur l'état de santé ou le risque pour la santé d'une personne (ex : croisement d'une mesure de poids avec l'âge, la taille, etc.) (données de santé « **par croisement »**) ;
- les données qui deviennent des données de santé en raison de l'utilisation qui est faite au plan médical, y compris dans le cadre d'une recherche en santé portant sur des échantillons biologiques humains (données de santé « **par destination** »).

La notion de donnée personnelle de santé doit être appréciée au cas par cas en fonction de la nature des données recueillies.

Les données génétiques sont également définies comme « les données à caractère personnel relatives aux caractéristiques génétiques héréditaires ou acquises d'une personne physique qui donnent des informations uniques sur la physiologie ou l'état de santé de cette personne physique et qui résultent, notamment, d'une analyse d'un échantillon biologique de la personne physique en question » (art. 4.13).

Du point de vue de la recherche en santé, deux types de données de santé semblent à distinguer au premier abord : les données produites à des fins premières de recherche, qu'il s'agisse de RIPH ou de NRIPH, et celles qui peuvent avoir un intérêt pour la recherche tout en étant

primitivement recueillies soit dans le cadre des soins, soit à des fins « médico-administratives » premières, aussi appelées « données de routine » (PMSIs, SNIIRAM/SNDS). Selon la loi, « sont exclus du cadre des traitements à caractère personnel dans le domaine de la santé les traitements (de données) entrant dans le cadre du soin individuel des personnes, ceux qui entrent dans le service des prestations d'assurance maladie obligatoire ou complémentaire ou ceux qui entrent dans le cadre de la gestion de l'information médicale dans les établissements de santé ». Mais dès lors que ces données sont réutilisées dans le cadre d'une recherche en santé, elles entrent dans le cadre des Traitements de données à caractère personnel dans le domaine de la santé (art. 64 et suivants de la LIL) et relèvent donc des formalités afférentes.

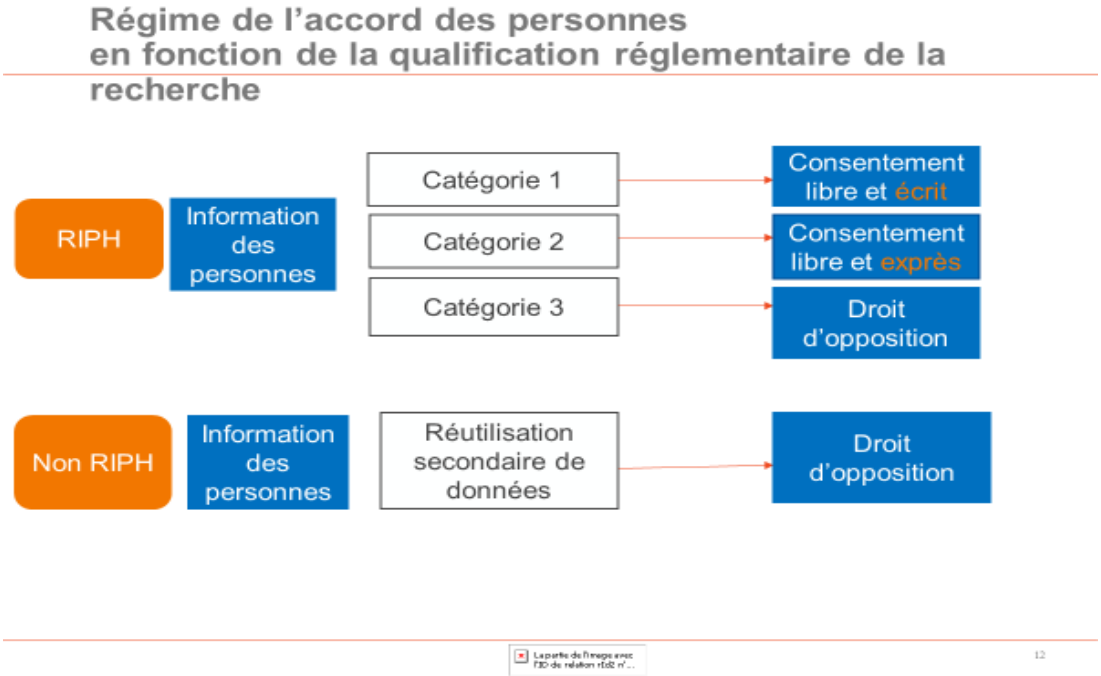
L'information médicale hospitalière ne se limite pas aux données appartenant au noyau du SNDS, celles des différents Programmes de Médicalisation du Système d'information. Il s'agit également des données de soin qui sont stockées dans les entrepôts de données de santé (EDS). Ces entrepôts contiennent des masses de données identifiantes très sensibles (de l'orthogénie à la psychiatrie en passant par les séquençages génomiques) recueillies par des personnels qui cumulent souvent fonctions de soin et de recherche, ce qui les amène à jongler en permanence sur les deux registres du soin personnalisé dans le cadre d'une relation interpersonnelle de confiance couverte par le secret médical et celui de la recherche qui mobilise ces données à d'autres fins que celles pour lesquelles elles ont été initialement recueillies sauf dans le cas de protocoles de recherche clinique relevant d'emblée de la RIPH (Amiel et Dosquet, 2021[xxviii]). Une vigilance particulière s'impose en matière de respect des droits des personnes, notamment celui d'être informées des traitements à base de réutilisation de leurs données de soin.

Pour tenter de simplifier, on peut raisonner en termes de « cycle de la donnée » en distinguant par commodité l'émergence de la donnée (« production initiale de données primaires ») et sa réutilisation (pour d'autres fins que celles pour lesquelles elle a été initialement collectée).

I.1.2 Cycle de la donnée, règles juridiques, information des personnes et confidentialité

Production initiale / émergence des données de santé primaires dans le cadre de la recherche et obligations d'information des personnes

La qualification réglementaire de la recherche a une incidence directe sur les formalités applicables au projet (dépôt de demande d'autorisation ou engagement de conformité à une méthodologie de référence) et sur la nature des droits reconnus aux personnes (*voir ci-dessous Régime de l'accord des personnes en fonction de la qualification réglementaire de la recherche*). Si certaines recherches dans le domaine de la santé sont dispensées de formalités ou peuvent être mises en œuvre sans délai par les responsables de traitement lorsqu'elles sont conformes à une méthodologie de référence, d'autres doivent être autorisées par la CNIL.



Quelle que soit la qualification de la recherche et le régime de formalités applicables, l'information des personnes est essentielle. Le RGPD garantit une plus grande transparence à l'égard des personnes concernées visant à leur conférer une meilleure maîtrise de leurs données personnelles.

En matière de recherche en santé, cette transparence est une garantie essentielle reconnue aux participants en contrepartie de la levée du secret professionnel. Elle leur permet de comprendre les objectifs de la recherche, les modalités de leur participation, la portée de l'accord qu'elles donnent et de maîtriser l'utilisation qui sera faite de leurs données. Elle conditionne et facilite l'exercice effectif de leurs droits. Elle permet d'instaurer une relation de confiance avec les chercheurs.

Par principe, l'information doit être délivrée individuellement à chaque personne participant à la recherche que les données soient recueillies auprès d'elle ou de tiers et spécifique à chaque projet (LIL, art. 58). Elle doit être réalisée pour chaque projet auquel le patient participe ou pour lequel les données du patient feront l'objet du traitement

L'information individuelle doit être doublée d'une information générale dans les lieux de soins qui transmettent des données à caractère personnel pour permettre des activités de recherche (affichage dans les locaux, mention dans le livret d'accueil, etc.)

Les conditions sont précisément encadrées dans le Code de la santé publique pour ce qui relève des données de recherche (Amiel, Dosquet et CEEI, 2021[xxix]). Dans le cas de recherche impliquant la personne humaine (RIPH), les études ne peuvent être légalement mises en œuvre qu'après un avis favorable (avis éthique) d'un comité de protection des personnes (CPP) et celui-ci examine avec une attention particulière les modalités de recueil du consentement des personnes, donc la clarté et la complétude de l'information qui leur sera donnée pour se déterminer. Le CEEI-IRB de l'Inserm fait de même dans son champ de compétences, càd pour les recherches non qualifiées RIPH, comme les comités d'éthique créés par les universités. Dans le cas de recherches dans le domaine de la santé non interventionnelles entrent des études qui ne requièrent pas le recueil du consentement mais celui de la manifestation d'une non-opposition. L'information des personnes est tout autant cadrée par les textes et les Méthodologies de Référence auxquelles peuvent se rattacher ces protocoles en définissant précisément les exigences à satisfaire. Selon la CNIL, « La méthodologie de référence MR-003 encadre les traitements comprenant des données de santé et présentant un caractère d'intérêt public, réalisés dans le cadre de recherches impliquant la personne humaine pour lesquelles la personne concernée ne s'oppose pas à participer après avoir été informée. Il s'agit plus précisément des recherches non interventionnelles et des essais cliniques de médicaments par grappe. L'information individuelle des patients est obligatoire. Le responsable de traitement s'engage à ne collecter que les données strictement nécessaires et pertinentes au regard des objectifs de la recherche. »

« La méthodologie de référence MR-004 encadre quant à elle les traitements de données à caractère personnel à des fins d'étude, évaluation ou recherche n'impliquant pas la personne

humaine. Il s'agit plus précisément des études ne répondant pas à la définition d'une recherche impliquant la personne humaine, en particulier les **études portant exclusivement sur la réutilisation de données**. » (La délibération n° 2018-155 du 3 mai 2018 de la CNIL précise « La recherche doit présenter un caractère d'intérêt public. Le responsable de traitement s'engage à ne collecter que les données strictement nécessaires et pertinentes au regard des objectifs de la recherche. »). Ses exigences sont claires et portent des obligations d'information collective et individuelle

Réutilisation et partage des données et information des personnes concernées

La réponse à certaines questions de recherche peut nécessiter, et ces usages se multiplient, la réutilisation de données (de recherche ou de routine), ce dont on ne peut que se réjouir en termes de valorisation du travail initial fourni pour leur « production ». La réutilisation des données ne change pas la qualification de la recherche. C'est le cas des cohortes, dont la succession des études entrent le plus souvent dans un cadre RIPH3. La réutilisation secondaire ponctuelle de données constitue un traitement de données personnelles distinct du traitement source. Ce nouveau traitement est soumis à des formalités propres et à une information spécifique des personnes concernées,

Le point essentiel étant que tout nouveau traitement répond à une nouvelle finalité qui nécessite dès lors une nouvelle procédure d'information des personnes. Ce point est très sensible et il nous a été fortement souligné par les membres du bureau de l'association Constances comme par les responsables scientifiques de la cohorte du même nom.

La réutilisation de données intervient également quand des données de soin sont utilisées secondairement à des fins de recherche. Elles relèvent souvent de la MR004 et cette dernière vise principalement à encadrer la réutilisation exclusive de ces données.

Il est indispensable de concevoir des modalités d'exercice des droits souples et dynamiques, susceptibles de permettre aux infrastructures de répondre aux enjeux d'une meilleure compréhension des mécanismes pathologiques par exemple, tout en garantissant une meilleure maîtrise par les personnes de leurs données, ce qui participe de la confiance indispensable des personnes concernées. Comme le souligne Georges Dagher au sujet des biobanques, « (. . .) *il est temps de repenser le rôle des personnes-sources plus largement en termes de participation et de contribution à la recherche (. . .). Le nouveau paradigme développé par l'utilisation des collections biologiques et visant à créer une ressource pour la recherche invite à une évolution du cadre réglementaire et éthique qui régit la question de la participation des patients aux projets de recherche* (. . .) » (Le Monde supplément

Sciences & Santé du mercredi 8 juillet 2015). Il faut saluer l'évolution de la doctrine de la CNIL sur ce point, illustrée par le projet de méthodologie de référence MR004 (*voir encadré « Information des personnes et respect des droits « informatique et libertés » dans le cadre d'une recherche sous MR004 portant exclusivement sur des données et/ou des échantillons biologiques*)

Chaque projet conforme à la MR 004 doit être enregistré dans un répertoire public tenu par la Plateforme des données de santé ou Health data hub et accessible sur son site internet.

« Information des personnes et respect des droits « informatique et libertés » dans le cadre d'une recherche sous MR004 portant exclusivement sur des données et/ou des échantillons biologiques

La MR004 admet que des données et/ou des échantillons biologiques puissent faire l'objet d'une réutilisation et que l'information puisse être considérée comme valablement délivrée dès lors que :

1. « Une information générale concernant les activités de recherche dans l'établissement doit être assurée auprès des personnes concernées (affichage dans les locaux, mention dans le livret d'accueil, etc).
2. A cette information générale s'ajoute l'information individuelle du patient inclus dans les recherches. Elle doit être réalisée pour chaque projet auquel le patient participe ou pour lequel les données du patient feront l'objet du traitement.
3. Des données et/ou des échantillons biologiques recueillis non spécifiquement pour la recherche peuvent faire l'objet d'une réutilisation sans qu'il soit procédé à une nouvelle information individuelle des personnes concernées :

. Lorsque la personne concernée dispose déjà des informations prévues aux articles 13 ou 14 du RGPD ; ceci pourrait, par exemple, concerner plusieurs projets de recherche, menés par un même responsable de traitement avec des finalités identiques, des catégories de données identiques et des destinataires identiques ;
. Ou lorsque l'information délivrée lors de la collecte des données et / ou des échantillons biologiques prévoit la possibilité de réutiliser les données et/ou les échantillons, et renvoie à un dispositif spécifique d'information auquel les personnes concernées pourront se reporter préalablement à la mise en œuvre de chaque nouveau traitement de données (par exemple : un site Internet sur lequel serait présenté chaque projet de recherche mené sur les données et/ou échantillons collectés dans le cadre de l'information initiale). »

Il peut s'agir d'un site internet, centralisant les informations relatives à l'ensemble des projets menés leurs caractéristiques et auquel les personnes pourront se reporter.
A défaut, une dérogation à l'obligation d'information sera toujours possible mais l'absence d'information des personnes rendra la recherche non éligible aux Méthodologies de référence et une demande d'autorisation sera nécessaire.

Recommandations :

Dans une logique de transparence, il est indispensable d'anticiper le devenir des données personnelles collectées dans la cadre d'un projet initial en mentionnant dans la note d'information, le partage futur des données et leur réutilisation (y compris dans des pays tiers) pour une ou plusieurs finalités de recherche en santé et en renvoyant vers un dispositif spécifique d'information, tel qu'un « portail de transparence », auquel les personnes pourront se reporter pour ne pas avoir à contacter directement et personnellement la personne avant la mise en œuvre d'un nouveau traitement.
L'information générale ne dispensera pas de l'information individuelle préalable spécifique à chaque nouveau projet de recherche qui nécessite la réutilisation secondaire de données déjà collectées mais si une documentation des projets sur ce site internet est prévue, les personnes concernées seront mises en mesure de s'informer si elles le souhaitent.
Si cela n'est pas fait, une dérogation à l'obligation d'information sera toujours possible mais :
- l'absence d'information des personnes rendra la recherche non éligible aux méthodologies de référence de la CNIL et une autorisation de la CNIL sera nécessaire ;
- le responsable du traitement devra donc effectuer une demande de dérogation à l'obligation d'information dans le dossier de demande d'autorisation qui devra être adressé à la CNIL qu'il devra solidement documenter en justifiant de l'impossibilité d'informer les personnes, efforts disproportionnés ou le fait d'informer rend impossible ou compromet gravement la réalisation de l'étude.

Réutilisation et partage des données entre plusieurs organismes de recherche

Les équipes INSERM peuvent être amenées à « partager » des données personnelles avec d'autres équipes de l'INSERM (le Responsable de traitement et le DPO restent les mêmes) mais également d'ailleurs, Pasteur par exemple (Responsable de traitement et DPO changent).

Lorsqu'on constitue une base de données dans le cadre d'un projet de recherche initial, des soins…, quel que soit sa qualification réglementaire, il faut se poser la question de la réutilisation des données contenues dans plusieurs projets de recherche (« entrepôt »), informer les personnes concernées de la réutilisation secondaire des données dans le cadre de projets de recherche scientifiques, comme indiqué précédemment, et recueillir le consentement explicite des personnes concernées à cette réutilisation.

En cas de consentement explicite non recueilli, le traitement relatif à la constitution de l'entrepôt doit faire l'objet d'une demande d'autorisation « santé » (hors recherche) et la finalité poursuivie doit revêtir un caractère d'intérêt public (un référentiel CNIL est en cours d'élaboration).

Dans tous les cas, une analyse d'impact sur la protection des données (AIPD) devra être réalisée.

S'il s'agit de données RIPH 1 ou 2, un consentement exprès de participation à la recherche est obligatoire et la possibilité de réutilisation des données peut être prévue par le biais d'une case à cocher supplémentaire. Dans une étude non interventionnelle qui ne nécessite « que » la manifestation de non-opposition, le recueil du consentement explicite pour la réutilisation secondaire des données à des fins de recherche est plus problématique.

La réutilisation secondaire ponctuelle de données constitue un traitement de données personnelles distinct du traitement source et, dans la mesure où ce nouveau traitement est soumis à des formalités propres et à une information spécifique des personnes concernées, l'INSERM, responsable du traitement « source », devient fournisseur de données pour le projet scientifique subséquent et il lui appartient de ne communiquer que les données strictement nécessaires au projet et de s'assurer que le tiers est « autorisé ». Il faut un contrat explicite avec les partenaires pour entériner une rupture de responsabilité (c'est l'équipe destinataire des données pour un autre traitement qui est responsable de ce traitement et donc des données qu'elle utilise à cette fin).

Sachant que si le partenaire demande des données de Elfe par exemple, ce sera un sous-ensemble précisément défini, aussi économe que possible (minimisation des données), désidentifié, et avec l'aval du comité scientifique de Elfe (« sensibiliser les guichets ») au regard de sa pertinence eu égard à la question de recherche posée et des conditions de sécurité et de protection des données personnelles (pas de possibilité de réidentification du fait de nouvelles possibilités de croisement). Dans cette chaine de « responsabilités successives », l'INSERM doit s'assurer que tout transfert se fait au profit d'un destinataire autorisé à les recevoir ou qui s'engage à respecter une MR, avant de pouvoir considérer que les données transférées sont alors sous la responsabilité du destinataire. Une gouvernance de l'accès aux données associant les responsables des bases sources doit ainsi être mise en place pour documenter la conformité du projet sur le plan scientifique et réglementaire.

Mais si le partenaire, pour les besoins de son protocole de recherche, a besoin d'administrer des questionnaires auprès des personnes qui ont participé à la recherche initiale, il ne peut le faire que s'il a accès à l'identité des personnes, ce qui n'est possible que par l'intermédiaire de l'équipe à l'origine du premier traitement qui seule a la possibilité de recontacter les personnes et de les informer sur les finalités du nouveau traitement. Et cette recherche secondaire devra elle-même être qualifiée et suivre le circuit ad hoc.

Comme l'écrivent des chercheurs néerlandais (Jacobs et Popma, 2019), « *Data governance should not end with sharing* ».

I.1.3 *Problèmes posés par l'inscription de collections de données constituées à des fins de recherche au « catalogue » du HDH.*

Dans un but de facilitation du partage de données de sources variées à des fins elles-mêmes variées, le HDH encourage les « producteurs de données » à les mettre à disposition sur la plateforme nationale des données de santé HDH. Cette invitation concerne les bases de données constituées par les chercheurs et placées sous la responsabilité de l'institution INSERM. Si le RGPD semble encadrer de manière satisfaisante (« *processor agreement* ») les relations habituelles entre « controller » (« *the responsible person / organization providing the means and determining the goals of data processing ») et « data processor » (« processing the data on behalf of the controller* »), la plupart des configurations de recherche n'entrent pas dans ce cadre et aboutissent à un transfert de responsabilité sur les données à celui qui les héberge / traite. Le problème nait du fait que le consentement a été donné à celui qui est à l'origine de la collecte alors que le nouveau dépositaire des données (le HDH dans notre cas) et celui qui va les traiter n'ont aucun lien personnel avec les sujets qui ont consenti à leur collecte pour un usage précis ni aucune connaissance de leur identité ; ils n'ont aucun moyen de les informer individuellement en vue d'une demande de consentement à une réutilisation de ces données, généralement pour une autre fin que celle pour laquelle elles ont été recueillies. Ceci peut être une source majeure de défiance de la part des personnes participant à la recherche et des protocoles de garantie des engagements initiaux doivent être prévus entre les responsables initiaux de la collecte et ceux à qui elles ont été transférées (Jacobs et Popma, 2019). Or, il importe, de ne pas compromettre la participation des participants aux cohortes, particulièrement les cohortes en population générale où la participation des personnes repose sur le volontariat. Et comme la durée de vie de ces collections dépasse généralement le temps dédié au projet initial par l'investigateur principal, le protocole doit couvrir l'ensemble du cycle de vie des données…

Selon Jacobs et Popma (2019), les bases légales pour la réutilisation de données partagées comportent 4 dimensions :

- garantir que l'usage délégué reste dans le cadre des engagements pris au moment de la collecte initiale ; l'utilisateur des données transférées a la possibilité effective de respecter les obligations du premier « controller » ;
- renforcer les exigences de protection des données, confidentialité et pseudonymisation en conformité avec les normes ISO ;
- protéger la propriété intellectuelle, y compris celle relative à l'ensemble des données dérivées (« derived from other data or from biosamples ») ;

- permettre à l'investigateur initial de pouvoir vérifier le respect de ces obligations avec une possibilité de révocation de l'agrément en cas de non-respect.

1.1.4 Préoccupations liées au respect du RGPD (hors hébergement) dans le cas des EDS

Selon la CNIL[xxx], « Les entrepôts de données sont créés principalement pour collecter et disposer de données massives (données relatives à la prise en charge médicale du patient, données sociodémographiques, données issues de précédentes recherches etc.). Ces données sont ensuite réutilisées, principalement à des fins d'études, de recherches et d'évaluations dans le domaine de la santé. Ces bases de données sont constituées pour une longue durée (plus de 10 ans en général) et l'objectif est d'obtenir un volume de données important. Elles peuvent être alimentées par de multiples sources (professionnels de santé, patients, pharmacie, établissements de santé, etc.). »

La CNIL a adopté depuis le début des travaux du groupe un Référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé. « Le référentiel permet aux organismes voulant mettre en œuvre un entrepôt de données conforme au référentiel de ne pas solliciter d'autorisation préalable auprès de la CNIL : après vérification de la conformité de son projet d'entrepôt par rapport au référentiel, l'organisme peut déclarer sa conformité. En interne, l'organisme responsable de ce traitement est tenu de documenter sa conformité au RGPD et au référentiel dans son registre des activités de traitement. Le référentiel entrepôt ne s'applique qu'aux entrepôts de données de santé reposant sur l'exercice d'une mission d'intérêt public, au sens de l'article 6.1.e du RGPD. »

Dans le contexte actuel de fonctionnement des établissements de santé, des doutes s'expriment quant à la réalité et l'effectivité de l'information des personnes dont les données sont collectées, la nature des utilisations qui peut en être faite et notamment celle de réutilisation à des fins pour lesquelles elles n'ont pas été collectées[xxxi]. Ceci vaut déjà pour les EDS indépendamment du HDH du fait du caractère sensible des données qu'ils renferment (un simple contact avec le système de santé est une information privée, mais il s'agit potentiellement de données plus intimes (vie sexuelle, conduites addictives, santé mentale, …) ou d'intérêt pour les assureurs (cancérologie), encore plus eu égard à la durée de stockage de certaines données qui peuvent avoir un intérêt différé que ce soit pour le soin ou la recherche, génétiques par exemple.

Cette préoccupation se renforce du fait des possibilités d'appariement des bases de données à un nombre de plus en plus important d'autres bases permis par le HDH à travers son catalogue. Ne serait-

ce que parce que dans ces conditions, des données anonymes ne le restent plus du fait que la réidentification des personnes est possible du fait du croisement possible d'informations très précises.

La situation des EDS semble parfois opposée à celle de dispositifs de recherche comme les cohortes : les personnes dont les données sont enregistrées dans les EDS, utilisatrices de services de santé, sont censées autoriser l'enregistrement de leurs données de soin au fil de l'eau sur la base d'une information éclairée qui vaut également pour la réutilisation de ces données à des fins de recherche. L'information semble minimale de fait (note de bas de page en petits caractères sur les documents « administratifs » (comptes-rendus ou convocations aux divers RV notamment) et celle sur la réutilisation doit faire l'objet d'une recherche active sur un site informatique plus ou moins explicite et facile à consulter. A l'opposé, dans une cohorte comme Constances [en italiques, extraits du CR d'audition], *« l'enjeu de protection des données et des droits des usagers est l'objet du « contrat moral » entre les volontaires et la cohorte, ce qui est à la base d'une relation de confiance (information et consentement sont d'ailleurs formalisés à toutes les étapes de la participation aux activités de recherche). Cette confiance va d'abord aux coordinateurs de la cohorte (qualité de la relation établie par des chercheurs respectés) et à leur appartenance institutionnelle (INSERM comme organisme public de recherche). Elle se fonde également sur la mise en place de procédures explicites d'information et de recueil du consentement à participer aux travaux (depuis l'acceptation initiale à participer après avoir été tiré au sort à partir de la base de données du RNIAM (Répertoire national interrégimes des bénéficiaires de l'assurance maladie) jusqu'au consentement demandé pour chacune des opérations de recherche : questionnaires annuels et spécifiques, inclusion dans la biobanque [son hébergement au Luxembourg soulève lui aussi quelques réticences de la part de certains volontaires]. Les webinaires organisés par les coordinateurs de la cohorte à destination des volontaires, celui sur le CASD tout particulièrement, sont appréciés*[xxxii]*. »*

Les responsables de cohortes sont très attentifs au respect du contrat de confiance qui les lie aux volontaires.

« Les données sont la propriété des volontaires, mais il est clair pour tous qu'elles sont recueillies à des fins de partage (« nos données nous appartiennent, on accepte de les partager, la question c'est avec qui » ; « sans chercheurs, nos données ne servent à rien » [association] ; « en aucun cas je ne considère que ce sont mes données, ce sont des données confiées » [chercheuse]) ».

« Tout projet de recherche repose sur une réutilisation des données qui suppose l'accord explicite des volontaires. Ceux-ci sont informés par la lettre de la cohorte et ils peuvent s'opposer individuellement à la réutilisation de leurs données pour un projet spécifique. De fait, les oppositions sont rares (quelques

dizaines à chaque fois, plutôt pour des projets du privé) mais elles suffisent à témoigner que l'information circule. »

Il semble que la philosophie qui préside à la mise en place d'entrepôts de données de santé puisse varier. Certains EDS ont intégré dès leur conception, au-delà du respect des droits de personnes, la nécessité de les considérer comme des partenaires. Le Ouest Data Hub par exemple, s'est constitué dans un « écosystème de la donnée » qui considère que ce n'est pas de la donnée que l'on partage mais de l'expertise autour de la donnée. Et cette expertise est clinique. Mais un tel écosystème suppose la confiance des personnes qui confient leurs données et cette dernière repose sur la transparence avec un effort constant d'information des personnes.

I.2 Qualité des données et qualité des projets de recherche

I.2.1 Qualité des données : données épidémiologiques vs données de routine

Rappelons que selon le RGPD, les données doivent être « exactes ». C'est la force des dispositifs de recueil de données ad hoc tels que les enquêtes épidémiologiques ou les cohortes, organisés pour recueillir des données de qualité maitrisée (même si « on peut faire de très belles choses avec le SNDS » selon les propos de chercheurs entendus lors d'une audition). Cette qualité résidant aussi dans la granularité plus fine que celle des données de routine. Cet avantage se payant de plusieurs manières : coût de fonctionnement de ces infrastructures, taille « limitée » (même si une cohorte comme Constances compte près de 220.000 volontaires) qui empêche l'étude de maladies rares, absence de représentation pédiatrique (pour Constances mais il y a Elfe), plus généralement biais de sélection liés au caractère volontaire de la participation.

Les responsables scientifiques de la cohorte Constances donnent des exemples précis qui permettent d'illustrer la supériorité des cohortes pour la production de connaissances épidémiologiques valides. L'intérêt de la granularité des données peut s'illustrer dans le cas des critères diagnostiques : Constances différencie d'emblée diabètes de type 1 et 2 ce que la connaissance d'un traitement par l'insuline, seule information présente dans le SNDS, ne suffit pas à distinguer ; il est alors possible de calibrer un algorithme de distinction des types de diabète avec les seules données du SNDS par rapport à cette référence diagnostique. Des travaux sont en cours dans le cadre du réseau ReDSiam pour évaluer la capacité du SNDS à identifier des problèmes de santé définis sur la base d'algorithmes utilisant les seules données du SNDS et les validant par le recours à des données

complémentaires issues de Constances, de registres de maladies ou d'autres sources dont l'information diagnostique peut servir de référence.

La précision et le détail des données offrent également des possibilités de prise en compte de facteurs de confusion et de contrôle des biais (de classement et de sélection pour partie) inatteignables avec les seules données de routine. Etudier la consommation de soins des personnes obèses par exemple, suppose de prendre en compte plusieurs facteurs de confusion : départager le rôle de la seule obésité suppose de connaitre le poids des personnes (absent du SNDS) et de tenir compte (pêle-mêle) de la présence d'un diabète, de son ancienneté et de complications éventuelles, de la consommation de tabac (ancienneté et quantité), … toutes précisions absentes du SNDS et qu'on ne retrouvera que dans Constances. Sans oublier de rappeler la possibilité de tenir compte finement du statut social des personnes (cf. infra). Dans certaines conditions, une maitrise partielle des biais de sélection est possible : Constance a des informations sur la cohorte des non-volontaires (n=400.000) ce qui permet d'évaluer l'importance de ses biais de sélection et de produire des données épidémiologiques redressées.

Un autre avantage d'une cohorte comme Constances est d'être une infrastructure ouverte qui permet la réalisation de projets de recherche mobilisant les données de la cohorte et éventuellement des données appariées (plus d'une centaine de projets déposés). Par comparaison, le HDH est un hébergeur qui n'est pas en capacité d'aider les chercheurs à utiliser les données des bases de son catalogue pour au moins deux raisons :

- Certains projets de recherche supposent des recueils de données spécifiques qui nécessitent de pouvoir entrer en contact avec les volontaires qui présentent un profil d'intérêt ; seul un dispositif comme Constances a cette possibilité d'identification nominative qui permet l'information et le recueil du consentement des personnes ;

- Les projets de recherche requièrent une connaissance fine des données très détaillées collectées ; seuls les chercheurs qui les ont générées ont cette connaissance et sont de ce fait en situation de guider d'autres chercheurs dans le choix des données pertinentes pour répondre à une question de recherche précise. Un exemple : de nombreuses variables peuvent servir à caractériser le statut social dans Constances, la(es)quelle(s) retenir ? Constances emploie en permanence 4 référents épidémiologistes en charge de guider les chercheurs dans le choix et le traitement des variables pertinentes en regard d'une problématique spécifique.

I.2.2 Qualité des projets de recherche

Les doutes sur la qualité des projets naissent du fait qu'ils seront conduits dans un environnement (HDH) qui n'a pas la recherche pour finalité première, mais des buts d'optimisation du fonctionnement du système de santé en tirant parti de la conjonction de possibilités immenses de stockage de données et du développement de méthodes de traitement innovantes grâce à la mobilisation de l'intelligence artificielle (IA). La finalité d'optimisation n'est guère discutable en soi (voire, ne pas mettre à profit les potentialités ne serait pas éthique), la perspective peut l'être (selon l'appréciation que l'on porte sur les différentes conceptions de « l'optimisation » ; les enjeux d'efficience seront-ils par exemple mis en regard des enjeux d'équité ?), indépendamment du fait qu'on s'illusionne probablement sur la portée des innovations attendues. Une interrogation peut se poser quant au périmètre de ce qu'on considère relever de la recherche. Un champ entier de la recherche sur les services, programmes et politiques de santé s'appuie sur la mobilisation des « banques de données médico-administratives » en toute connaissance des limites de validité des données qu'elles contiennent (censément compensées par leur massivité et la puissance des algorithmes de traitement). Un autre domaine est celui du développement des méthodes d'IA pour l'analyse des données et l'aide à la décision. L'IA ne compensera jamais le différentiel de validité entre données de routine (celles des BDMA ne sont pas recueillies pour la recherche) et données générées par des chercheurs pour des finalités de recherche épidémiologique (qu'elle soit descriptive, analytique ou évaluative). La constitution (sous-représentation des chercheurs) et les modalités de fonctionnement (charge de travail) du CEESRES interrogent à cet égard par comparaison avec le fonctionnement des comités qui l'ont précédé CCTIRS puis CEREES).

II Des pistes de réflexion et d'action pour une utilisation éthique et responsable des données de santé « massives »

Les lignes qui suivent ne font qu'esquisser des pistes de réflexion à approfondir. Elles s'appuient sur le contenu des premières auditions menées pour commencer à circonscrire le sujet. Les pistes proposées ne représentent donc pas un programme structuré mais de premiers éléments à enrichir par d'autres lectures et auditions, avec les responsables du HDH, dans la dimension

européenne notamment, et différentes équipes de recherche actives dans le champ (celui de « l'intelligence artificielle » ou celui de « l'internet des objets » notamment).

Garanties de sécurité en matière de conservation des données

La réflexion du CEI a démarré à partir d'une profonde inquiétude née de la décision de confier l'hébergement du SNDS à une société privée étrangère, de droit américain US par surcroit. Puisque la réflexion des pouvoirs publics s'oriente désormais vers une reconquête d'un certain niveau de souveraineté, il semble d'autant plus important d'étudier des possibilités d'hébergement nationales (ou européennes). A côté de sociétés privées comme OVH (d'autant plus que sa crédibilité a été profondément endommagée par l'incendie de certains de ses centres de données à Strasbourg) ou Thalès, l'existence d'une solution publique comme le CASD (Centre d'accès sécurisé aux données) semble mériter d'être rappelée. Ce d'autant qu'une autre dimension de la sécurité mérite considération : la question de la centralisation versus la possibilité de traitement de bases de données distribuées (au lieu que les données se déplacent en masse vers les centres où s'effectuent conjointement stockage et traitements, ce sont les logiciels de traitement qui se déplacent sur les lieux de stockage des données ; la taille du « pot de miel » est réduite d'autant – il y en a plusieurs- et les risques de fuites également). Parmi les avantages du CASD, le fait qu'il héberge déjà des données extrêmement sensibles d'opérateurs publics (le service des impôts ayant des exigences de sécurité particulièrement élevées), données qui peuvent présenter un intérêt scientifique pour la recherche en santé par ailleurs (sur les inégalités sociales notamment). Mais le CASD a également développé une procédure complète de sécurité (physique et logique) qui semble particulièrement adaptée à l'appariement de données de recherche avec des données de routine : la création de bulles sécurisées où cohabitent les seules données nécessaires à un projet et les logiciels de traitement pertinents (y compris d'IA), hermétiquement closes et uniquement accessibles depuis des « SD-Box » sécurisées, verrouillées et scellées, avec authentification par carte à puce et biométrie. Certaines équipes INSERM y font d'ailleurs appel, comme Constances ou le Centre Pierre Louis. Parmi les autres intérêts du CASD, sa participation au projet International Data Access Network et le développement d'une procédure originale de certification de la qualité des publications par les revues scientifiques, en partenariat avec l'agence de certification CASCAD (bulles sécurisées dédiées pour l'accès de l'auditeur, de l'agence de certification Cascad, aux données sources des publications scientifiques).

La question économique et industrielle se pose au niveau local également. Si certains établissements, comme l'AP-HP, font le choix du logiciel libre, il est rarement complet et une première contrainte pour le développement des EDS (entrepôts de données de santé) est la maitrise du

partenariat industriel. Selon un de nos informateurs-clés, « il y a une économie autour des plateformes » et « tant mieux si ce sont des entreprises françaises qui aident à développer la Recherche & Développement ». L'enjeu selon lui, et il est national, est de ne pas rater la convergence avec une dimension financière lourde. La dimension humaine est majeure également ; il est difficile de recruter au prix du marché et la question se pose de savoir comment étoffer les équipes (des Centres de Données Cliniques dans l'organisation grand-ouest) et recruter dans le secteur public des data scientists de haut niveau dans les conditions de rémunération du secteur public.

Respect du droit des personnes à une information honnête et accessible

Si la loi « Informatique et Libertés » fête ses 40 ans et que les principes posés, pour l'essentiel inchangés, sont connus et respectés dans la pratique de la recherche, l'entrée en application du RGPD a profondément modifié l'approche de la protection en s'invitant dans la gouvernance et l'organisation des acteurs qui doivent documenter la conformité et son respect ; respect encore largement perfectible par l'ensemble des intervenants en santé. Beaucoup reste à faire pour que les données soient toujours « traitées de manière licite, loyale et transparente au regard de la personne concernée », « adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées », « exactes » ou encore « traitées de façon à garantir une sécurité appropriée des données à caractère personnel, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de mesures techniques ou organisationnelles appropriées » et pour que l'organisation et les procédures qui en attestent soient en place.

Nos auditions montrent que beaucoup de chemin reste à parcourir pour garantir une information loyale des personnes garante d'un consentement éclairé. Ceci se pose dans le cadre primaire des protocoles de recherche. Ceci se pose surtout vis-à-vis de la réutilisation des données, qu'elles soient recueillies dans le cadre sécurisé de la recherche ou surtout dans le cadre du soin. Dans ce dernier cas, les données recueillies au fil du colloque singulier sont souvent de nature très sensible ; leur potentielle réutilisation à des fins de recherche n'est pas systématiquement envisagée ni l'information à ce sujet donnée. Sachant que toute réutilisation est censée faire l'objet d'une nouvelle procédure d'information spécifique de la finalité scientifique poursuivie. Des modalités d'information systématique et de recueil effectif d'un consentement spécifique restent à mettre en place, notamment dans le cas des EDS. La MR004 donne des indications très précises. Les modalités d'information générale comportent affichage, Charte du patient, … et des modalités spécifiques individualisées doivent les compléter. La contrainte de temps ne saurait être minimisée et des

procédures innovantes voient le jour dont les possibilités de généralisation restent à évaluer avant une possible généralisation. Il s'agit d'un chantier lourd mais qui conditionne la confiance du public dans la conduite d'une recherche respectueuse des droits des personnes et la continuité de sa participation.

Reconnaissance d'un partenariat entre enquêtés et enquêteurs dans la production de connaissances scientifiques

Comme d'autres, les responsables du HDH nous ont dit vouloir promouvoir une « culture de la donnée ». Ceci semble essentiel et il nous semble qu'elle doit prendre en compte la dimension spécifique des données de recherche. Contrairement aux « données de routine », les données de recherche sont le fruit d'un double travail : celui des personnes qui s'engagent et se prêtent à un travail d'investigation, pour contribuer au progrès des connaissances, sur une base volontaire (elles y consacrent un temps qui ne saurait être ignoré) ; celui des chercheurs qui ont conçu et mis en œuvre des dispositifs de collecte (majoritairement par questionnaire lors des études de cohortes) et de traitement des données avec un souci permanent de validité à toutes les étapes du cycle de vie des données. L'ensemble repose sur un contrat de confiance réciproque qui se construit dans un partenariat exigeant au long cours, confiance extrêmement fragile toutefois et potentiellement menacée par l'introduction de tiers qui n'ont ni l'histoire ni la culture de cette relation, comme ceci pourrait se produire si des dispositifs de recherche étaient inscrits sans précautions au catalogue du HDH. Ce travail appelle par ailleurs des mécanismes de valorisation qui brisent son invisibilité et le reconnaissent à sa juste valeur quand ces données versent dans un pot commun indifférencié où elles sont agrégées à des données de routine. Ce besoin de reconnaissance est pluriel : au minimum symbolique (reconnaitre l'effort de participation à la production d'un bien commun) et financière (aucun bien n'est gratuit et les données ont un coût qui doit être compensé pour que leur qualité puisse se maintenir à un niveau élevé).

Développement d'une dimension participative de la recherche épidémiologique

Les responsables scientifiques de la cohorte Constances, tout comme les membres de l'association du même nom, donnent un exemple de relation contractuelle entre volontaires et chercheurs. Conscience des chercheurs de travailler avec des « données confiées », volonté des volontaires de pouvoir décider de participer ou pas aux différents projets en fonction de l'intérêt qu'ils en comprennent. Ceci dépasse le seul respect du droit au consentement éclairé et relève d'un véritable partenariat de recherche dans un cadre public protégé des intérêts privés (c'est ce que disent les

volontaires). Ceci suppose la mise en place dans la durée de mécanismes de partage d'information et de consultation sophistiqués où la place du numérique est centrale.

Le responsable du « Ouest Data Hub » explique l'importance de cette dimension participative dans la constitution et le fonctionnement en routine des entrepôts de données de santé également. On ne produit pas des données mais de l'expertise et sans cette expertise et la transparence de cette expertise il ne peut pas y avoir de confiance.

Garanties de validité scientifique des projets de recherche

Un observateur éminent du fonctionnement des comités d'expertise attire l'attention sur les contraintes de temps et de technicité qui pèsent sur l'évaluation des projets de recherche. Un équilibre délicat doit être trouvé entre « intérêt public » et rigueur scientifique (qui devrait aller de pair). Ceci repose sur un examen contradictoire par les pairs et la construction d'une culture commune d'examen des dossiers facilitant leur expertise collective (anticipation qui facilite la gestion des situations d'urgence par ailleurs). Ceci suppose au minimum une présence suffisante de chercheurs dans les comités dédiés (nombre, temps dédié et délais d'expertise compatibles avec une instruction rigoureuse).

La garantie scientifique des travaux de recherche suppose également une connaissance fine des données (pertinence en fonction de l'objectif de production de connaissance, limites de validité) et donc une association étroite des chercheurs aux travaux qui mobilisent les données dont ils ont l'expertise. De la qualité de cette association dépend notamment la prise en compte des facteurs de confusion et la maitrise des multiples biais dont doivent se prémunir au mieux toutes les recherches épidémiologiques. Les responsables de la cohorte Constances y insistent particulièrement, tout comme les partenaires du Ouest Data Hub (« jamais seul face aux données ! »).

De l'expérience rennaise, la dématérialisation du dossier patient à l'origine de la construction d'un EDS réussit si elle est portée par l'idée maitresse que les données collectées à des fins de soin peuvent être amenées à resservir, à des fins de recherche notamment. Dans cette approche, l'outil technique est certes important mais c'est autour de l'expertise médicale de traitement de la donnée clinique qu'il faut le développer (l'organisation autour de centres de données cliniques dans chaque établissement d'un réseau a valeur d'exemple). Les travaux de recherche consistent à étudier plutôt les questions qui se posent et d'examiner si les données pour y répondre sont disponibles. Ce modèle est l'inverse de celui des « data brockers » qui déploient d'abord des outils pour capter de la donnée pour la mettre « à disposition » (à titre onéreux) ensuite.

Renforcer l'acculturation et l'appui aux chercheurs

Nous revenons ici sur l'impératif de respect des droits et libertés des personnes participant à la recherche et ses conditions d'effectivité.

Il n'est pas simple pour des chercheurs, surtout quand ils sont cliniciens, de prendre conscience et de respecter en permanence les exigences d'une gestion de l'information clinique selon deux modes, celui de la relation de soin où le recueil de données individuelles nominatives va de soi, ces données étant par ailleurs exclues du cadre réglementaire des traitements à caractère personnel dans le domaine de la santé, et celui de la recherche, où la protection de l'anonymat doit être farouche, l'information des personnes soignées devant être permanente et accessible afin de garantir dans les deux cas un consentement ou une manifestation de non-opposition au recueil et à la réutilisation des données éclairés. Un effort général de pédagogie semble nécessaire et un dispositif de formation adapté reste à mettre en place.

Mais les exigences de respect du RGPD et des lois françaises dans la recherche sont complexes. Si la rédaction des procédures est parfois simple, elle nécessite le plus souvent des appréciations subtiles de la situation vis-à-vis des exigences de la CNIL et des arbitrages qui doivent être gérables entre les contraintes légales ou réglementaires et leurs possibilités de respect (information de personnes perdues de vue, durée de conservation et modalités d'actualisation périodique). Le CEEI-IRB fait sa part pour ce qui concerne son champ de compétences ; la DPO de même en innovant fortement pendant la période COVID pour accélérer l'instruction des demandes d'autorisation avec la CNIL. Un besoin de renforcement important de leurs moyens apparait clairement, de même que celui des instances locales d'appui à la recherche, ces instances locales étant de plus en plus amenées à étendre leur champ de compétence au-delà de la seule recherche clinique pour laquelle elles ont été initialement mises en place. Il s'agit d'un enjeu majeur pour la concrétisation des promesses de recherche éthique et responsable à l'INSERM et chez ses partenaires.

Remerciements

Les animateurs du groupe de réflexion du CEI tiennent à remercier l'ensemble des membres du groupe et au-delà, l'ensemble des membres du CEI, en premier lieu son président, Hervé Chneiweiss. La contribution de la présidente du CEEI-IRB de l'Inserm, Christine Dosquet, a été également particulièrement éminente. Le groupe a pu bénéficier de l'appui de l'ancienne déléguée à la protection des données (DPO) de l'INSERM, Frédérique Lesaulnier. Enfin, Catherine Bourgain a fait bénéficier le groupe de la réflexion d'un cercle de chercheurs en sciences sociales issus de l'EHESS.

Les remerciements vont avant tout aux personnes qui ont accepté d'être auditionnées pour instruire le groupe. La note ici présentée tente de traduire fidèlement leur propos mais ne saurait les engager en aucune manière. La responsabilité de son contenu est exclusivement celle de ses auteurs.

Ont contribué successivement jusqu'à l'écriture de cette note d'étape :

- Professeure Catherine Quantin, PU-PH en biostatistiques et informatique médicale, spécialiste d'information médicale au CHU de Dijon
- Dr Grégoire Rey, directeur du CépiDC de l'INSERM (Centre d'épidémiologie sur les causes médicales de décès)
- Mme Frédérique Lesaulnier, data protection officer de l'INSERM jusqu'à l'automne 2021
- Pr Jean-Louis Serre, président du CCTIRS puis du CEREES
- Mr Kamel Gadouche, directeur du CASD
- Dr Marie Zins, Pr Marcel Goldberg, responsables scientifiques de la cohorte Constances
- Mmes Florence Ghioldi, présidente, Frédérique Anne, vice-présidente, et Martine Dréneau-Fénerol, secrétaire générale de l'association Constances
- Mr Gérard Raymond, vice-président du Health Data Hub, Mme Caroline Guillot, directrice adjointe des relations associations et citoyens, Mr Alexandre Romainville, direction des relations associations et citoyens
- Mme Laure Maillant, Directrice du pôle Innovation et Données à la DSI de l'APHP
- Pr Marc Cuggia, PU-PH en biostatistiques et informatique médicale à Rennes. Membre de l'UMR 1099 Laboratoire du traitement de l'image et du signal INSERM – Université Rennes 1), Responsable de l'Equipe Données Massives en Santé. Coordonnateur du LabCom LITIS et du réseau interrégional des centres de données cliniques du Grand Ouest (GCS HUGO).

[i] La notice poursuit : « Il désigne les responsables de traitement, définit leur rôle et leurs missions. Il modifie en outre la composition de la liste des organismes, établissements, et services bénéficiant d'accès permanents aux données du système national des données de santé en raison des missions de service public qu'ils exercent. Il précise les règles applicables à cet accès permanent. Il prévoit les modalités d'exercice des droits des personnes concernées et notamment les conditions d'information des personnes auxquelles les données se rapportent ».

[ii] Zins M, Cuggia M, Goldberg M. Les données de santé en France : abondantes mais complexes. Méd Sci (Paris) 2021 ; 37 : 179-84.

[iii] Chevreul K, Delpierre C, Dourgnon P, et coll. Les données de santé, un patrimoine commun qui doit servir à améliorer le bien-être de tous. Le Monde Idées, 30 octobre 2020, p27.

[iv] Chneiweiss H. Big data et santé : questions éthiques p200-201 in Big Data à l'échelle de la société. Rémi Mosseri ed. Presses du CNRS 2016.

[v] Délibération n° 2021-067 du 7 juin 2021 portant avis sur le projet de décret portant application du II de l'article 1er de la loi n° 2021-689 du 31 mai 2021 relative à la gestion de la sortie de crise sanitaire (demande d'avis n° 21010600).

[vi] Décret n° 2021-848 du 29 juin 2021 relatif au traitement de données à caractère personnel dénommé « système national des données de santé ». JORF n°0150 du 30 juin 2021.

« Publics concernés : personnes dont les données sont recueillies à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi médical ou médico-social ou d'une enquête dans le domaine de la santé et qui alimentent le système national des données de santé, organismes publics et privés ayant vocation à mener des projets à des fins de recherche, d'étude et d'évaluation dans le domaine de la santé.

Objet : modalités de mise en œuvre du système national des données de santé.

Entrée en vigueur : le texte entre en vigueur le lendemain de sa publication.

Notice : le décret prévoit les modalités de gouvernance et de fonctionnement du système national des données de santé dont le périmètre est étendu à de nouvelles bases de données. Il désigne les responsables de traitement, définit leur rôle et leurs missions. Il modifie en outre la composition de la liste des organismes, établissements, et services bénéficiant d'accès permanents aux données du système national des données de santé en raison des missions de service public qu'ils exercent. Il précise les règles applicables à cet accès permanent. Il prévoit les modalités d'exercice des droits des personnes concernées et notamment les conditions d'information des personnes auxquelles les données se rapportent.

Références : les dispositions du décret sont prises en application de la loi n° 2019-774 du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé

[vii] « Par ailleurs, l'essentiel de ces données a vocation à être intégré progressivement dans un « SNDS centralisé » composé d'une base principale (comprenant à ce jour le « SNDS historique » et pouvant être dans l'avenir enrichi) et d'une base catalogue incluant d'autres bases de données considérées comme pertinentes pour les acteurs de la recherche. Cette intégration impliquera, quant à elle, la migration des données. Initialement envisagé comme un système décentralisé, la Commission relève que le choix du ministère s'oriente finalement vers une centralisation des données du SNDS. Elle prend acte que ce projet de décret vise à amorcer cette centralisation des données auprès de la CNAM et de la PDS et à encadrer uniquement la mise en œuvre de ce « SNDS centralisé ».

[viii] Voire, l'existence à l'étranger de contrats de mise à disposition de données des services publics de santé à des sociétés privées. Lemke C. Ma santé, mes données. Premier parallèle, 2021, 171p.

[ix] « Nvidia met en service son nouveau supercalculateur au Royaume-Uni, le plus puissant du pays, avec cinq partenaires dans le domaine de la santé : les laboratoires AstraZeneca et GSK, la Guy's and St Thomas' NHS Foundation Trust, le King's College London, et Oxford Nanopore. Doté d'une puissance de calcul de 8 pétaflops, il sera utilisé pour différents projets de recherche, dont le développement d'une compréhension plus approfondie des maladies du cerveau comme la démence. Mais aussi pour renforcer l'emploi de l'IA pour concevoir de nouveaux médicaments et améliorer la précision de la recherche de variations génétiques causant des maladies chez l'homme. "Cambridge-1 donnera aux chercheurs la capacité (…) de débloquer des indices sur les maladies et les traitements à une échelle et à une vitesse auparavant impossibles", déclare Jensen Huang, fondateur et PDG de Nvidia. AstraZeneca souhaite de son côté accélérer ses travaux sur l'utilisation de l'intelligence artificielle dans la pathologie numérique [?], qui implique la capture, la gestion, l'analyse et l'interprétation d'informations numériques. Tandis que GSK espère mieux prédire la santé humaine et développer de meilleurs médicaments, plus aptes à afficher des résultats positifs lors des essais cliniques. » («Nvidia démarre son supercalculateur au Royaume-Uni, avec des projets en santé ». L'Usine Nouvelle – 7 juillet 2021, rapporté dans Pharmaceutiques, 8 juillet 2021)

[x] APMNews 19 juillet 2021 : « Création de "l'Alliance française des données en vie réelle" pour dresser un pont entre les industriels et le Health Data Hub ». « Lancée mardi à l'initiative du cabinet de conseil Kynapse et du think tank "AI for Health", l'Alliance française des données en vie réelle va accueillir "5 ou 6 laboratoires pharmaceutiques" à partir du mois de septembre pour "faciliter et accélérer les projets de recherche sur les données de vie réelle" et favoriser les collaborations entre les industriels et le Health Data Hub (HDH), a expliqué jeudi après-midi à APMnews Stéphane Messika, CEO de Kynapse. »

[xi] « Le laboratoire AstraZeneca, la filiale digitale de La Poste, Docaposte, et la société Impact Healthcare, spécialisée dans le conseil en innovation, lancent Agoria Santé : une plateforme de collecte et d'analyse des données de santé au service d'une meilleure prise en charge thérapeutique des patients. "L'objectif est de fournir un cadre juridique, éthique et sécurisé aux acteurs de la santé afin de leur permettre d'aller plus vite dans la recherche", explique au Figaro Olivier Nataf, à la tête de la filiale française d'AstraZeneca. Le catalogue de données et de services proposés devrait s'enrichir au fil du temps, et des partenariats noués autour de thématiques communes. Cinq laboratoires et une dizaine d'autres acteurs du secteur (hôpitaux, universités…) sont déjà en discussion pour rejoindre l'initiative. Agoria fonctionnera aussi comme une plateforme marchande, avec des services payants aux hôpitaux, aux universitaires ou aux laboratoires. Ses fondateurs, qui ne communiquent pas le montant de leurs investissements ni les tarifs des services, espèrent ainsi rentabiliser la plateforme. Les nouveaux membres du consortium devront s'acquitter d'un ticket d'entrée. Et les "utilisateurs" paieront en fonction des services utilisés. Selon Olivier Vallet, PDG de Docaposte, l'ambition est aussi "d'accélérer l'usage du digital et de l'intelligence artificielle pour faire de la France un leader. La crise sanitaire n'a fait qu'accélérer la prise de conscience". (AstraZeneca, Docaposte et Impact Healthcare s'unissent pour l'accès aux données de santé. Le Figaro – 18 juin 2021. Cité par Pharmaceutiques du 6 juillet 2021).

[xii] Itw de Frédéric Dufaux, directeur général adjoint de Docaposte en charge de la santé (TechmedInfo 5 juillet 2021).

Docaposte renforce son empreinte dans la santé digitale avec le lancement d'Agoria Santé, en partenariat avec le laboratoire AstraZeneca et le cabinet Impact Healthcare. Une nouvelle plateforme dédiée aux entreprises pour réaliser et sécuriser leurs projets de recherche sur les données de santé. Quel rôle joue aujourd'hui Docaposte dans la transition numérique en santé ?

Docaposte est hébergeur de données de santé (HDS), avec la particularité d'être certifié sur les six couches, ce qui permet d'intégrer une dimension applicative à notre offre. Notre stratégie s'écrit autour de deux grands axes: structurer, collecter et analyser les données au service des industriels, dont les medtech et les laboratoires pharmaceutiques, tout en sécurisant les process, dont l'utilisation des algorithmes. Il ne s'agit pas de juger de leur pertinence scientifique, mais de s'assurer que le traitement de la donnée répond dans sa forme aux exigences réglementaires, et qu'il sera auditable. C'est notre valeur ajoutée par rapport à un hébergeur plus classique. Notre deuxième axe est d'intervenir comme opérateur de données de santé pour faciliter les échanges et l'interconnectivité des systèmes entre opérateurs. L'exemple type est notre travail sur le dossier pharmaceutique (DP) pour l'Ordre des pharmaciens, ou encore auprès du groupe d'hôpitaux privés Elsan, que nous accompagnons sur son assistant virtuel, Adel.

[xiii] Une société internationale leader en France dans les logiciels de gestion de pharmacies d'officine, Iqvia, a attiré l'attention des médias et de la CNIL : « 17/05/2021 · Le fonctionnement de l'entrepôt de données de la société IQVIA autorisé en 2018 a été mis en cause dans l'émission Cash Investigation qui sera diffusée le 20 mai prochain. La CNIL précise qu'à ce jour, elle n'a pas reçu de plainte relative au fonctionnement de cet entrepôt mais annonce, au regard des éléments portés à la connaissance du public, qu'elle diligentera des contrôles. »

[xiv] Il faut se méfier alors d'une ressemblance trompeuse avec le raisonnement inductif mis en œuvre dans les sciences sociales. Si le sociologue « construit son objet » à partir de l'observation, 1) cette observation est conduite de manière méthodique, le sociologue devant « garder la maitrise de ses questionnements en se tenant à distance des « prénotions » du sens commun ou des polémiques de l'actualité » ; 2) (« en partant du principe wébérien que les individus sont les « atomes élémentaires » d'une société, l'un des buts de la sociologie est d'analyser les relations entre eux » (…) en prenant en compte « l'ensemble des critères qui jouent un rôle dans les activités et lient les individus entre eux ». « Il n'existe aucune théorie universelle qui permettrait d'expliquer la manière dont s'articulent [les] variables … on ne peut se contenter d'aligner les variables en présupposant qu'elles ont le même poids, travers fréquent chez les statisticiens ». Extrait de Race et sciences sociales, Essai sur les usages publics d'une catégorie, Beaud S et Noiriel G, Agone Paris 2021 [pp 188-9].

[xv] Surtout à l'heure où le Health data Hub lance un appel à manifestation d'intérêt autour des algorithmes de ciblage, dans le cadre du projet BOAS (Bibliothèque ouverte d'algorithmes en santé).

[xvi] Eric Sadin. L'intelligence artificielle ou l'enjeu du siècle. Ed L'échappée, Paris, 2021, 298p.

[xvii] Propos recueillis par Laure Belot dans le cadre d'un entretien publié dans la rubrique Sciences et Médecine du journal Le Monde daté du 28 octobre 2020.

[xviii] La délibération de la CNIL en date du 30 juin 2021 le souligne « La Commission relève toutefois qu'un organisme responsable de traitement d'une base source alimentant la base principale ou la base catalogue pourra continuer à mettre à disposition les données de la base source auprès d'autres responsables de traitement (par exemple, l'Agence technique de l'information sur l'hospitalisation mettant à disposition les données du PMSI auprès d'autres responsables de traitement ou un centre hospitalier universitaire mettant à disposition les données d'un entrepôt hospitalier auprès d'une entreprise spécialisée dans l'intelligence artificielle). Elle relève que cette mise à disposition sera régie par les dispositions du CSP (interdiction de poursuite des finalités interdites, respect du référentiel de sécurité du SNDS, etc.). Elle prend acte des précisions apportées par le ministère selon lesquelles l'organisme est responsable du traitement de sa base source tant qu'il traite les données et jusqu'à ce que les données soient traitées pour alimenter le « SNDS centralisé ». »

[xix] Mais le champ de la réflexion s'arrête là, le groupe n'adresse pas la question de la sécurité des applications mobiles, pourtant problématique. Voir par exemple Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis. Grundy Q, Chiu K, Held F, Continella A, Bero L, Holz R. BMJ 2019;364:l920

[xx] Soulignés dans la délibération CNIL du 30 juin 2021 : « Le ministère a confirmé que la PDS disposera d'une copie de la base principale, pour répondre efficacement aux demandes et notamment pour la réalisation d'appariements ad hoc entre la base principale et le catalogue. Sans remettre en cause cette nécessité opérationnelle, la Commission s'inquiète toutefois de la duplication d'une base comportant, par nature, des données sensibles couvrant l'ensemble de la population. En effet, cette duplication implique de transférer régulièrement un grand volume de données entre la CNAM et la PDS, ainsi que de partager des identifiants pseudonymisés ; en outre, la Commission rappelle que la PDS ne dispose pas – contrairement à la CNAM – de ses propres centres de données et fait appel à un prestataire dans un centre de données mutualisé avec plusieurs clients. Elle rappelle que ces différentes opérations augmentent mécaniquement la surface d'attaque et les risques de violations sur ces données. »

[xxi] Goldberg M, Zins M. La plateforme « Health Data Hub » pose des questions de sécurité majeures. Le Monde Idées, 30 octobre 2020, p27.

[xxii] Banck A. RGPD : la protection des données à caractère personnel. 19 fiches pour réussir et maintenir votre conformité. 3ème édition, Gualino, Paris 2020, 79p.

[xxiii] Selon la CNI, « La mise en œuvre de traitements de données à caractère personnel intervenant dans le cadre de la recherche s'effectue sous la responsabilité du responsable de traitement, et/ou chez des tiers agissant pour son compte. Le responsable de traitement doit effectuer une analyse d'impact relative à la protection des données, qui doit couvrir en particulier les risques sur les droits et libertés des personnes concernées. Il met en œuvre les mesures techniques et organisationnelles appropriées afin de garantir un niveau de sécurité adapté aux risques identifiés. Une seule et même analyse peut porter sur un ensemble d'opérations de traitement similaires. Le responsable de traitement doit mettre en œuvre et contrôler l'application d'une politique de sécurité et de confidentialité en application de la méthodologie de référence.

[xxiv] L'article 4 du RGP les définit ainsi : « Constitue une 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 » (en pratique identification directe, indirecte ou par recoupement) ».

[xxv] Le même article 4 du RGPD précise : « Constitue un traitement de données à caractère personnel toute opération […] portant sur de telles données […] notamment la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction ».

[xxvi] « Digital data are taking on an ever more central role in biomedicine today. But these data are increasingly generated outside the traditional spaces of the medical system, as individuals go about their daily lives interacting with consumer devices. Moreover, the technological tools needed to produce, store and analyze these data increasingly lie beyond the remit of traditional medical scientists. In other words, the health data ecosystem is expanding, to include new types of data, new methods for capturing and analyzing them, and new stakeholders. Pressing questions emerge concerning privacy, informed consent, the commodification of personal health data, and the drawing of new power asymmetries between data subjects and data controllers, the public and the private sector. At the same time, concepts and values that previously acted as normative anchor points, such as "solidarity", the "public" or the "common good", are destabilized, re-conceptualized, and mobilized in new ways. This special theme addresses the question of how the expansion and decentralization of the health data ecosystem disrupts existing norms and frameworks of data ethics and data governance, and what kinds of re-thinking of ethics and governance this solicits. The collection of articles and commentaries provides a

combination of conceptual and practice-based reflection. » Présentation de la rubrique Health Data Ecosystem de la revue Big Data and Society par ses éditeurs, Tamar Sharon, Associate Professor, Radboud University et Federica Lucivero, Senior Researcher, University of Oxford (accédé en ligne le 18 juin 2021).

[xxvii] « Digital data are taking on an ever more central role in biomedicine today. But these data are increasingly generated outside the traditional spaces of the medical system, as individuals go about their daily lives interacting with consumer devices. Moreover, the technological tools needed to produce, store and analyze these data increasingly lie beyond the remit of traditional medical scientists. In other words, the health data ecosystem is expanding, to include new types of data, new methods for capturing and analyzing them, and new stakeholders. Pressing questions emerge concerning privacy, informed consent, the commodification of personal health data, and the drawing of new power asymmetries between data subjects and data controllers, the public and the private sector. At the same time, concepts and values that previously acted as normative anchor points, such as "solidarity", the "public" or the "common good", are destabilized, re-conceptualized, and mobilized in new ways. This special theme addresses the question of how the expansion and decentralization of the health data ecosystem disrupts existing norms and frameworks of data ethics and data governance, and what kinds of re-thinking of ethics and governance this solicits. The collection of articles and commentaries provides a combination of conceptual and practice-based reflection. » Présentation de la rubrique Health Data Ecosystem de la revue Big Data and Society par ses éditeurs, Tamar Sharon, Associate Professor, Radboud University et Federica Lucivero, Senior Researcher, University of Oxford (accédé en ligne le 18 juin 2021).

[xxviii] Amiel P, Dosquet C, Comité d'Evaluation Ethique de l'Inserm (CEEI). Guide de qualification des recherches en santé. Inserm, 2021.

[xxix] Voir aussi Astruc A, Jouannin A, Lootvoet E, Bonnet T, Chevallier F. Les données à caractère personnel : quelles formalités réglementaires pour les travaux de recherche en médecine générale ? Exercer 2021 ; n°172 : 178 – 184.

[xxx] Traitements de données de santé : comment faire la distinction entre un entrepôt et une recherche et quelles conséquences ? 28 novembre 2019 Consulté sur le site de la CNIL le 2 octobre 2021.

[xxxi] Délibération de la CNIL du 30 juin 2021 : « La Commission relève que, malgré l'ampleur du traitement, tant en termes de sensibilité que de volume des données, le projet de décret ne prévoit pas d'information individuelle des personnes concernées. Par ailleurs, prenant acte que l'information sera presque exclusivement réalisée de façon dématérialisée (sites web, compte Ameli), la Commission demande au ministère de réfléchir à des modalités d'information supplémentaires alternatives (campagnes d'affichage ou d'information dans les media, mise à disposition des notes d'information dans les caisses primaires d'assurance maladie, transmission d'une note d'information complète en cas de demande des personnes concernées, etc.). Quant aux 30 % des assurés ne disposant pas d'un compte Ameli, la Commission demande qu'une information individuelle complète leur soit délivrée par voie postale, par exemple, à l'occasion de l'envoi d'un relevé de remboursement. »

[xxxii] Extrait du compte-rendu de l'audition de représentants de l'association Constances et des responsables scientifiques de la cohorte le 12 avril 2021.