Bénéfices et risques de l'utilisation des données de santé à des fins de recherche

Ségolène Aymé, R Choquet, L Devillers, M Gilard, M Kelly-Irving, A Livartowski, B Lukacs, D Polton

Conseil scientifique consultatif du Health Data Hub

Bénéfices et risques de l'utilisation des données de santé à des fins de recherche

Rapport - octobre 2023

Auteurs :

Ségolène Aymé, (présidente du Conseil scientifique consultatif du Health Data Hub, Inserm) ; Rémy Choquet (Roche SAS) ; Laurence Devillers (Sorbonne Université) ; Martine Gilard (CHU de Brest) ; Michelle Kelly-Irving (Inserm) ; Alain Livartowski, (Unicancer Paris) ; Bertrand Lukacs (vice-président du Conseil scientifique consultatif du Health Data Hub, Académie nationale de chirurgie) ; Dominique Polton (HCAAM).

Pour citer cet article :

S. Aymé, R. Choquet, L. Devillers, M. Gilard, M. Kelly-Irving, A. Livartowski, B. Lukacs, D. Polton, "Bénéfices et risques de l'utilisation des données de santé à des fins de recherche", Conseil scientifique consultatif du Health Data Hub, octobre 2023.

Remerciements :

Le groupe de travail remercie Pierre Lombrail (Professeur honoraire de santé publique, Laboratoire Educations et Promotion de la Santé, Université Sorbonne Paris Nord ; membre du Comité Ethique de l'Inserm) pour ses commentaires précieux qui ont enrichi la réflexion et aidé à la rédaction de ce document.

Résumé exécutif de l'article

La disponibilité de nombreuses données de santé, maintenant que la numérisation du système de santé devient une réalité dans les pays développés, suscite des attentes de la part de tous les acteurs, parfois surestimées, mais aussi des craintes, parfois exagérées. Il est proposé une revue des bénéfices que l'on peut légitimement attendre de l'exploitation des données disponibles, et des maléfices déjà documentés ou que l'on peut légitimement anticiper, puis de regarder de quel côté penche la balance bénéfices/risques. Nous proposons que la balance bénéfice/risque soit dorénavant le support des décisions d'autorisation pour l'utilisation des données de santé à des fins de recherche, alors que les critères actuels privilégient la protection absolue de la vie privée, ce qui entrave ou retarde la réalisation de nombreuses études de grande pertinence.

Les bénéfices attendus de l'accessibilité de ces données sont nombreux.
(1) Elles facilitent la surveillance épidémiologique des grandes pathologies, base de beaucoup de décisions en santé publique ;
(2) elles permettent de mesurer l'efficacité des soins et de détecter des situations locales ou régionales s'écartant significativement de la moyenne ;
(3) de même pour le suivi de l'évolution des pratiques dans le temps et leur impact ;
(4) elles permettent de documenter les effets positifs et négatifs des innovations qui sont introduites par l'évolution des pratiques, qu'elles soient médicamenteuses, techniques ou organisationnelles ;
(5) elles informent sur l'accès des patients aux interventions de santé : on peut calculer des taux et délais, et les croiser avec des variables géographiques, des données sociales et économiques, et ainsi identifier des inégalités, ce qui fonctionne ou non, et sur quels territoires doivent porter les efforts d'amélioration en priorité ;
(6) elles contribuent directement à l'amélioration des connaissances, étape indispensable pour améliorer les pratiques et concevoir des politiques publiques efficaces et protectrices.
(7) elles sont indispensables pour développer les usages de l'intelligence artificielle grâce à de nouveaux algorithmes de bonne qualité car entraînés et testés sur des populations représentatives, ce qui n'est pas toujours le cas actuellement ;
(8) elles permettent d'évaluer l'efficacité des campagnes de prévention ou de dépistage, et l'observance des recommandations de bonnes pratiques.

Mais l'usage de ces données par de multiples acteurs n'est pas sans **risques**.
(1) le risque de révélation de données personnelles est un risque majeur pour les données nominatives mais très faible pour les données pseudonymisées et nul pour les données anonymisées ;
(2) le risque d'utilisation non contrôlée, au-delà de l'information des personnes, dépend du respect des règles d'utilisation par les opérateurs. Il est très surveillé ;
(3) le risque de vol de données et d'espionnage par des industriels ou des états est souvent cité par les opposants, il est pourtant extrêmement réduit car les données non nominatives n'ont pas de valeur marchande et ce type de vol ne s'est encore jamais produit ;
(4) le risque de violation de la propriété intellectuelle et de préjudice financier par obstacle à la valorisation est souvent cité mais non documenté ;
(5) le risque de surveillance normative des professionnels de santé est craint par ces professionnels, alors que cette surveillance est légitime du point de vue des financeurs du système de santé et les tensions peuvent être apaisées par le dialogue.

Les risques potentiels pour les citoyens et la collectivité sont très minimes si les mesures de protection qui ont été adoptées sont efficacement mises en œuvre, et aucun des risques n'est de nature à faire courir un danger à fort impact (de santé), ni individuel ni collectif. En

revanche, le bénéfice de l'utilisation de ces données peut être important pour la collectivité (ex. pertinence des soins, prévention des scandales sanitaires, évaluation des innovations) et pour l'individu (ex. développement de nouveaux traitements) et avoir des impacts sur les individus eux-mêmes. La balance bénéfice/risque est donc très en faveur des bénéfices.

Reste à **lever les freins** à la mise en œuvre de l'ouverture des collections de données et à diffuser dans toutes les couches de la société la culture du bien commun que représentent les données de santé. Ces freins sont nombreux et comprennent :

(1) l'interdiction d'utiliser des outils américains pour stocker des données alors qu'aucun cloud européen n'est disponible, et qu'il n'y a jamais eu de plaintes dans le secteur de la santé contre des industriels américains qui ne respecteraient pas le RGPD dans le domaine de la santé, ni de documentation de tentatives de réidentification de données anonymisées de recherche en santé ;

(2) la difficulté d'utilisation du NIR pour apparier les bases de données que l'on souhaite croiser, alors que cela est possible dans beaucoup de pays européens, soumis à la même réglementation européenne que la France ;

(3) des délais d'attente se chiffrant en année pour accéder à la base principale du SNDS alors que la file d'attente se réduirait si une copie de la base principale du SNDS était au catalogue du HDH ;

(4) un écosystème trop complexe pour des raisons historiques, qui a besoin d'être simplifié pour aller vers un guichet unique, comme la directive européenne en préparation le préconise ;

(5) une communication grand public ne portant que sur les risques et non sur les bénéfices, alors que les citoyens bien informés sont en faveur de l'usage de leurs données pour la recherche.

Cela aurait un impact très important sur la faisabilité de nombreuses études. Un audit des entraves à la recherche dans le système actuel serait utile pour convaincre les réticents.

Introduction

La disponibilité de nombreuses données de santé, maintenant que la numérisation du système de santé se généralise dans le monde, suscite des attentes de la part de tous les acteurs, parfois surestimées, mais aussi des craintes, parfois exagérées. Il est donc opportun de faire une revue de la réalité des bénéfices que l'on peut attendre de l'exploitation des données disponibles, et des risques déjà documentés ou que l'on peut anticiper, et regarder de quel côté penche la balance bénéfices/risques.

Cette réflexion s'inscrit dans plusieurs politiques publiques adoptées récemment, en particulier la loi "organisation et transformation du système de santé" (1), et un plan "Pour une science ouverte" (2) qui est une transposition d'une directive européenne. Des arguments théoriques plaident en faveur d'une meilleure utilisation des données disponibles pour la décision en santé publique et pour une accélération de la recherche et développement au service de la qualité de vie des citoyens (3). Toutes les enquêtes réalisées à ce jour montrent que les citoyens attendent que les données de santé servent à améliorer notre santé publique et notre système de soin (4).

Le partage des données signifie leur circulation entre de nombreux acteurs, à partir de multiples bases de données dispersées, peu connues de l'extérieur et difficiles à utiliser par ceux qui ne les ont pas constituées. Les procédures d'accès aux données de santé peuvent être perçues comme complexes, en raison même de la sensibilité de ces données. Elles obéissent à des gouvernances différentes, parfois discrétionnaires. Les outils et les compétences nécessaires pour traiter la donnée de manière sécurisée, comme cela s'impose légitimement, sont coûteux et bien souvent inaccessibles, en particulier pour des petites équipes de recherche, des institutions dont la recherche n'est pas l'activité principale, ou des start-ups.

La France est déjà très engagée vers une mutualisation des données de santé disponibles, notamment au travers de la loi santé de 2019, laquelle a élargi le système national des données de santé et créé une plateforme fédératrice d'accès aux données : le Health Data Hub (5). Ces choix suscitent des questionnements, parfois des tensions, souvent liés à la complexité du sujet, d'où l'importance d'apporter des éclairages complémentaires nécessaires à une bonne compréhension du bien-fondé du chemin emprunté (6,7).

Le Comité Consultatif National d'Ethique (CCNE), dans son avis n°130 (8), contribue à apporter des réponses par une analyse couvrant tous les aspects éthiques de la numérisation des données, bien au-delà du présent article qui se consacre uniquement aux questions éthiques que pose l'utilisation des données de santé pour la recherche, et non pour le soin.

Le 9 mai 2023, le Comité consultatif national d'éthique (CCNE) et le Comité national pilote d'éthique du numérique (CNPEN) ont également rendu public un avis commun relatif aux plateformes de données de santé (9). Au total, les deux comités formulent 21 recommandations : en matière de partage des données, de leur anonymisation, mais aussi de souveraineté ou encore d'impact environnemental. Ainsi, le CCNE et le CNPEN portent notamment leur attention sur le contrôle des données de santé issues des "cliniques privées, des EHPAD, et plus récemment des plateformes de biologie largement acquis par des groupes financiers relevant de fonds d'investissements internationaux".

Cet article a pour objectif de présenter les implications de la politique de science ouverte pour les données de santé, et d'ainsi contribuer à fournir les informations utiles à un débat informé, en particulier pour apprécier la balance bénéfices/risques, pilier d'une décision éclairée.

Il discute aussi l'écosystème institutionnel et réglementaire, garant de la protection de la vie privée et de la confiance des citoyens, écosystème qui ne doit pas entraver la recherche, en privilégiant les droits individuels sur les droits collectifs et en ne considérant que les risques potentiels sans souci des bénéfices attendus.

Il a aussi pour objectif de rappeler que l'écosystème de la recherche est international et que les initiatives nationales doivent être en harmonie avec les autres systèmes de régulation de la recherche, pour ne pas entraver les partenariats transnationaux, indispensables à la recherche et à l'innovation.

La finalité de ce travail est de proposer une revue des bénéfices et des risques générés par la réutilisation de données de santé à des fins de recherche, et de suggérer des évolutions de l'écosystème de mise à disposition des données de santé pour la recherche, pour mieux servir la santé publique et la qualité du système de soin.

1. Les données de santé et leur organisation

Une donnée de santé fait référence à toute information relative à l'état de santé, physique ou mental, d'une personne (10). Il peut s'agir d'informations relatives à l'identification d'une personne à des fins de santé (numéro, symbole, etc.) ; d'informations relatives à des tests ou examens, y compris des données génétiques et biologiques ; d'informations relatives aux maladies, symptômes, traitements, handicaps, antécédents, de données sur le style de vie si elles ont un impact potentiel sur la santé ; ou d'informations relatives à la consommation de soins ou de services, en particulier via des dispositifs médicaux ou du processus de remboursement des soins.

Il s'agit donc de toutes les informations produites lors de consultations, d'hospitalisations ou d'investigations radiologiques ou biologiques, d'enquêtes (qu'elles soient informatisées ou non) ou d'enregistrements administratifs. De plus, des données qui ne paraissent pas être des données de santé, sont considérées comme telles si, du fait de leur croisement avec d'autres données, elles permettent de tirer une conclusion sur l'état de santé ou le risque pour la santé d'une personne : croisement d'une mesure de poids avec d'autres données (nombre de pas, mesure des apports caloriques, par exemple). De même des données peuvent devenir des données de santé en raison de leur destination, c'est-à-dire de leur utilisation à des fins médicales.

Seules les données numérisées sont prises en compte dans cet article, car ce sont les seules à pouvoir être exploitées directement à des fins de recherche.

Nous allons maintenant passer en revue les différentes sources de données de santé utilisables à des fins de recherche.

1.1. Données administratives

De nombreuses administrations disposent de données de santé pour assurer le remboursement des soins, la gestion du système de santé et la qualité des services aux personnes. La principale collection de données médico-administratives est la base principale du système national des données de santé (SNDS), créé en 2016. Cette base rassemble et met à disposition des informations de santé collectées par des organismes publics et pseudonymisées, c'est-à-dire dans lesquelles le numéro de sécurité sociale (ou NIR, Numéro d'inscription au répertoire) est remplacé par un pseudonyme généré de façon aléatoire. Elle comprend des informations issues de trois bases de données : le Système National d'Information Inter Régimes de l'Assurance Maladie (SNIIRAM) ; le Programme de

Médicalisation des Systèmes d'Information (PMSI) qui centralise les informations de facturation des établissements de santé ; la base des causes médicales de décès (CépiDC), alimentée par les données des certificats de décès. Un même individu dispose du même pseudonyme dans les trois bases et les informations le concernant peuvent être rapprochées. La loi de 2016 prévoyait d'enrichir cette base principale du SNDS par l'ajout de deux autres sources de données médico-administratives : les données "médico-sociales" des maisons départementales des personnes handicapées, et un échantillon représentatif des données de remboursement par les organismes complémentaires.

D'autres sources de données viennent également enrichir le patrimoine français, comme les données de passages aux urgences de Santé Publique France ou les données de déclaration des maladies à déclaration obligatoire.

Il existe aussi d'autres sources administratives de données de grand intérêt pour la recherche, qui dépendent d'autres organismes, comme par exemple l'INSEE. Parmi ces dernières, on retrouve notamment l'échantillon démographique permanent (EDP) qui est un panel socio-démographique de grande taille créé en 1967. Il rassemble des informations de nature essentiellement démographique issues de cinq sources : les bulletins d'état civil de naissance, de mariage et de décès depuis 1968 ; les données issues des recensements entre 1968 et 1999 puis des enquêtes annuelles de recensement depuis 2004 ; les données du fichier électoral depuis 1967 ; les données du panel " tous salariés " depuis 1967 ainsi que les données socio-fiscales depuis 2011. L'EDP est accessible pour les chercheurs ou organismes extérieurs au service statistique public via le Centre d'accès sécurisé aux données après avis du Comité du secret statistique.

1.2. Données de soin et entrepôts hospitaliers

Les établissements de soin enregistrent aussi l'activité de soin relative à chaque usager. C'est le dossier patient qui est de plus en plus informatisé pour permettre une meilleure traçabilité des soins ainsi qu'un meilleur partage d'information entre professionnels intervenant dans les soins. Généralement, les données recueillies sont réduites à l'essentiel pour ne pas alourdir la charge de travail des professionnels de santé. Ces données n'en sont pas moins intéressantes pour la recherche bien qu'elles ne soient pas toujours adaptées à cette activité. Jusqu'à présent, les hôpitaux n'étaient pas organisés pour les exploiter à des fins de recherche mais cela est en train de changer : des entrepôts hospitaliers de données sont en construction ou en déploiement sur tout le territoire (11) et permettront l'utilisation secondaire de ces données. Ces entrepôts contiendront en effet la partie des données de soin considérée comme exploitable à des fins de recherche et d'intérêt. Celles-ci seront pseudonymisées.

De même, les données de soin par les praticiens libéraux sont une source précieuse d'information pour connaître l'état de santé de la population, maintenant que de plus en plus de cabinets sont informatisés. Cependant il existe actuellement une grande variété de logiciels de gestion de cabinet. Un consortium à l'initiative du collège national des généralistes enseignants (CNGE), le projet P4DP (12), a pour ambition de créer en trois ans un entrepôt de données de médecine de ville au plan national, ce qui donnera enfin accès à des données de vie réelle en médecine de ville.

L'avantage des données administratives et des données de soin est qu'elles sont collectées en routine et de façon exhaustive. Elles couvrent, pour la plupart, tous les citoyens bénéficiant d'une assurance santé, soit plus de 95% de la population. Les utiliser en recherche permet de les valoriser à moindre coût, pour un bénéfice collectif important. Il y a cependant un investissement à consentir pour les rendre accessibles aux chercheurs, en raison des contraintes réglementaires et techniques qui s'imposent pour protéger ces

données sensibles. Comme nous le verrons, il faut, en particulier, que les personnes concernées ne se soient pas opposées à l'usage de leurs données pour la recherche et que les données elles-mêmes existent dans un format répondant à des standards nationaux ou internationaux, aussi bien techniques que sémantiques.

1.3. Cohortes, registres et collections de données pour la recherche

Les chercheurs dans le domaine de la santé ont constitué au fil du temps de grandes bases de données pour disposer des données nécessaires pour valider ou invalider des questions de recherche. Ces bases de données ont des formats variés, allant du registre collectant des évènements de santé de manière transversale, aux cohortes (13) ayant l'ambition de suivre de façon longitudinale les personnes ayant accepté de confier leurs données pour une finalité de recherche. Les données des essais cliniques pour évaluer les bénéfices et les risques de nouveaux traitements, sont également des collections de données de grand intérêt. Une autre catégorie est constituée des données d'enquête pour explorer des dimensions médicales mais aussi sociales ou psychologiques, dans des populations spécifiques volontaires ou en population générale. Depuis peu, des données sont également générées par des objets connectés, données destinées à la surveillance, mais pouvant être réutilisées pour d'autres types de recherche. Toutes ces données sont formatées pour la recherche et collectées dans un cadre qui oblige à une information individuelle des personnes, ce qui n'est pas toujours possible. Cela rend ces données directement utilisables pour des finalités différentes de la finalité d'origine, à condition que les personnes en soient informées et qu'elles puissent exercer leurs droits. Néanmoins, ces collections de données sont coûteuses à maintenir et limitées dans leur périmètre géographique. Ainsi, une analyse récemment développée par le Haut Conseil de la Santé Publique pointe ces limites et propose des solutions, dont notamment l'appariement de ces collections avec la base principale du SNDS (14).

1.4. Données environnementales et sociales pouvant influer sur l'état de santé

Les données environnementales utiles pour qualifier et décrire l'effet des facteurs environnementaux sur la santé sont diverses, et peuvent concerner, par exemple, les sources de nuisance, les contaminants des milieux, ou l'occupation des sols. Parmi les données environnementales pertinentes sur le champ de la santé-environnement peuvent par exemple être citées les données de l'Ineris sur la qualité de l'air (Géod'air, PREV'AIR, Cartothèque) ou collectées par les agences agréées de la qualité de l'air (AASQA), les données de qualité de l'eau remontées des ARS (SISE-Eaux), les données météorologiques de Météo France, les données des sites et sols pollués (Infosols), les cartographies du bruit du Cerema (plaMADE) ou encore les données de vente des produits phytosanitaires gérées par l'OFB (BNV-D).

Les environnements humains sont marqués par une hiérarchie sociale, donnant lieu au gradient social de santé, c'est-à-dire la relation entre le niveau social et l'état de santé. En effet, le plus on s'élève dans cette hiérarchie sociale, meilleur est l'état de santé, et vice versa[1]. Afin de bien comprendre comment les inégalités sociales de santé sont construites tout au long de la vie et selon des catégories sociales différentes (genre, position

[1] Sur la période 2012-2016, l'espérance de vie à la naissance pour les 5 % les plus riches de la population était de 84,4 ans pour les hommes et de 88,3 ans pour les femmes, contre 71,7 ans et 80 ans respectivement parmi les 5 % les plus modestes, soit un écart socioéconomique de treize ans pour les hommes et de huit ans pour les femmes (Blanpain, 2018) (15). De plus, les catégories sociales moins favorisées font face à une « double peine » mêlant à la fois une durée de vie plus courte et une qualité de vie moins bonne que les catégories sociales aisées. Les populations les plus défavorisées en France vivent aussi dans les lieux où ils sont plus exposés aux multiples pollutions environnementales qui sont néfastes pour la santé humaine (16).

socioéconomique, groupe ethnique, lieu de vie, etc.), il est primordial que les données sociales de qualité soient chaînées et analysées en lien avec les données de santé diverses (diagnostique, incidence, suivi, traitement, biomarqueurs). Cela permettra de fournir des données probantes pour la recherche interventionnelle visant à prévenir et réduire les inégalités sociales de santé.

1.5. Le SNDS vise à représenter la diversité des données de santé

Initialement composé des données médico-administratives telles que les feuilles de soin, la facturation hospitalière et les causes médicales de décès, le SNDS a été élargi par la loi relative à l'organisation et la transformation du système de santé du 24 juillet 2019 à toutes les données de santé qui bénéficient d'un financement de la solidarité nationale dans le but d'élargir le patrimoine des données disponibles et de contribuer ainsi à une utilisation plus large des données.

Le SNDS comprend donc désormais des données de registres, de cohortes de recherche, d'entrepôts de données hospitalières, etc.

Le HDH est chargé par la loi de réunir, organiser et mettre à disposition les données du SNDS élargi et de promouvoir l'innovation dans l'utilisation des données de santé. Le décret publié en juin 2021 dispose que le HDH est co-responsable de traitement de la base principale avec la CNAM et responsable de traitement du catalogue du SNDS. Ces deux sous-ensembles du SNDS présentent un intérêt majeur pour l'écosystème.

Ainsi, la base principale désigne la réunion de données, couvrant l'ensemble de la population, en provenance de l'Assurance Maladie (base SNIIRAM), des établissements (base PMSI), des causes médicales de décès (base du CépiDC de l'Inserm), des données relatives au handicap (en provenance des MDPH - données de la CNSA) en cible. Elle a vocation à être enrichie en continu avec l'intégration d'autres bases de données nationales et notamment, dans le cadre de ce premier arrêté, des bases relatives à l'épidémie : Vaccin-covid et SI-DEP.

Le catalogue désigne quant à lui une collection de bases de données, non figée. En effet, construit itérativement, il permet de s'adapter aux enjeux et besoins de l'écosystème, et son contenu est fixé par arrêté, tout comme les flux de la base principale. La première version de cet arrêté a été publiée le 12 mai 2022 et prévoit que 10 bases de données constituent la première version du catalogue du SNDS. Parmi celles-ci, la banque nationale de données maladies rares (BNDMR) centralisant les dossiers patients informatisés créés par les centres de référence ; la base de données relative à l'exploitation des données de passages aux urgences, dénommée "OSCOUR" ; ou encore La base de données relative aux données de surveillance de 33 maladies à déclaration obligatoire permettant de prévenir les risques d'épidémie, dénommée "Maladies déclaration obligatoire" (MDO). Une deuxième version du catalogue du SNDS est par ailleurs en cours de construction.

2. Les bénéfices attendus de leur utilisation

Il existe de nombreuses collections de données déjà collectées qui peuvent être reliées entre elles pour un enrichissement mutuel des données et qui sont d'un grand intérêt pour faire progresser nos connaissances. Pourtant les bénéfices de cette mutualisation des données pour la recherche ne sont pas clairement identifiés par beaucoup de citoyens comme de

professionnels de santé et de décideurs, car la communication vers tous ces publics pointe essentiellement les risques et non les bénéfices.

Une revue des bénéfices potentiels est donc présentée ci-après, et résumée dans le tableau en annexe :

2.1. Bénéfices généraux : masse critique, représentativité des populations étudiées, complémentarité

L'intérêt de l'utilisation et de la réutilisation de données déjà disponibles est évident. C'est, *a minima*, une valorisation des coûts investis pour la collecte des données, et *a maxima*, une occasion d'améliorer les connaissances et les pratiques de façon significative. Cet avantage est d'autant plus grand que l'événement que l'on cherche à analyser est rare, car il faut alors de grands effectifs, jamais disponibles au niveau d'une seule équipe de recherche ou d'un établissement de santé. Le regroupement de bases de données permet de disposer des effectifs nécessaires pour conclure sur des évènements rares et de comparer des évènements entre régions géographiques ou entre groupes sociaux. Il donne de la puissance aux études et de la robustesse aux résultats.

Le croisement de bases de données permet d'enrichir les données de chacune. Ainsi, les données sur les décès en France permettent de savoir si des personnes figurant dans des bases consacrées à une maladie sont encore en vie ou non, information qu'il est impossible d'obtenir par d'autres moyens. Les variations de fréquence de maladies entre des régions peuvent être expliquées par des corrélations avec des expositions environnementales venant de bases de données sur les implantations industrielles, les décharges, les incinérateurs ou les niveaux de pollution de l'air ou de l'eau, par exemple. Ce type de travaux est impossible à réaliser autrement que par le croisement de bases de données.

La bonne gestion de données sensibles et l'usage optimal de bases de données complexes conçues par d'autres, ne sont pas aisés à mettre en œuvre car ils requièrent des compétences rares pour lesquelles peu de personnes ont été formées. Il y a du sens à partager les compétences existantes car beaucoup d'institutions n'arriveront pas à recruter les professionnels des données dont ils ont besoin. Cela requiert aussi une infrastructure technique offrant de grandes capacités de calcul et un haut niveau de sécurité, qui est trop complexe et coûteuse pour être disponible dans toutes les institutions constituant des collections de données. Des infrastructures partagées par de multiples acteurs sont la seule solution d'un point de vue financier, de compétence et de performance. Ceci ne signifie pas regrouper les bases de données dans un seul entrepôt national, mais regrouper une copie de ces bases si elles doivent être chaînées entre elles, pour être mises à la disposition des chercheurs. Il peut y avoir plusieurs entrepôts spécialisés établis selon les besoins de communautés particulières. Il faut simplement que toutes ces bases de données puissent être identifiables grâce à leur présence dans des entrepôts de données pour la recherche, et interopérables entre elles par le respect des normes sémantiques et techniques.

2.2. Bénéfices pour la surveillance épidémiologique

La surveillance épidémiologique est une activité de santé publique qui a pour objet de collecter, de façon continue, des informations sur des événements de santé, de les analyser pour construire des indicateurs chiffrés et de les cartographier, puis de les diffuser, afin de produire une aide aux décideurs dans le domaine de la santé. Essentiellement développée depuis les années 1950, elle est devenue au fil des décennies, avec la succession des crises sanitaires, un outil indispensable à l'élaboration et la conduite de toutes politiques de santé.

Traditionnellement, des bases de données spécifiques ont été construites pour générer l'information recherchée, essentiellement des registres de morbidité d'intérêt pour Santé

Publique France ou l'INCA par exemple. Ces registres ont de nombreuses limites, car ils ne couvrent qu'une fraction de la population, ce qui limite leur puissance statistique et ils sont très coûteux à maintenir car ils nécessitent la participation de nombreux professionnels pour en assurer la qualité et l'exhaustivité (17). De plus, ils ne sont pas faits pour détecter des signaux faibles et nouveaux, puisqu'ils ne collectent que des données qui ont déjà été identifiées comme étant d'intérêt.

Les registres et cohortes, dans leur mission de surveillance épidémiologique, ont donc des limites mais peuvent devenir d'excellents outils s'ils sont couplés aux données de soin. Ils peuvent bénéficier directement des bases de données administratives pour collecter des données sur le parcours de soin des personnes incluses, et vérifier le statut vital de celles-ci. On peut aussi valider le degré d'exhaustivité des enregistrements ce qui rend plus robustes les travaux de recherche produits. Les registres et cohortes peuvent être aussi très utiles pour permettre la validation d'hypothèses générées par les données de soin et la construction d'enquêtes complémentaires en cas d'alerte. La surveillance à partir des données de soins permet de détecter des fluctuations d'incidence d'évènements de santé, et de les analyser par territoire et par sous-populations pertinentes, permettant ainsi l'identification d'inégalités sociales de santé et dans l'accès aux soins, ou de différences d'exposition à des facteurs de risque locaux. Cela répond au souhait des citoyens de surveillance des taux d'incidence de pathologies comme les maladies infectieuses, les cancers, les maladies respiratoires ou cardiovasculaires, les anomalies du développement embryofœtal, ou toute autre pathologie dont les déterminants sont largement environnementaux. Cette surveillance est techniquement faisable à un coût raisonnable pour une efficacité très grande, sans aucun effet délétère potentiel. La surveillance épidémiologique a besoin des registres et cohortes mais aussi des données de soins. Seules les données de vie réelle que sont les données de soin, permettent de valider les conclusions des études de recherche qui sont toujours menées sur des populations particulières et dans des conditions contrôlées, parfois éloignées de la réalité. Nos connaissances seraient optimales si les différentes sources de données étaient interopérables et connectées.

2.3. Bénéfices pour l'amélioration des pratiques

Les données de santé collectées dans le cadre du soin, pour aider à la prise en charge individuelle des malades ont aussi un très grand intérêt pour améliorer les prises en charge d'un point de vue collectif. Elles permettent de mesurer l'efficacité des soins et de détecter des situations locales ou régionales s'écartant significativement de la moyenne. Elles permettent de suivre l'évolution des pratiques dans le temps et leurs impacts. Par la base principale du SNDS, il est possible d'analyser le parcours de soins de certaines pathologies : ce retour d'informations, reflet de la vie réelle, est très important pour les professionnels car il met en évidence des réalités qui interrogent sur la pertinence des prises en charge. Ainsi, la base Observapur (18) qui analyse le parcours de soins des hommes traités pour troubles de la miction montre que le recours à la chirurgie, toute chose étant égale par ailleurs, varie de 1 à 3 selon des régions. On a pu mettre en évidence de telles variations géographiques pour de nombreux actes chirurgicaux : chirurgie de la fracture de hanche, ablation des amygdales, appendicectomie, césarienne, chirurgie de l'obésité… (19). Les données du SNDS peuvent aussi permettre de mettre en évidence des écarts entre les recommandations médicales et les pratiques, et ainsi orienter des actions d'information et de formation des professionnels.

2.4. Bénéfices pour l'évaluation des innovations

Les données permettent de suivre les effets positifs et négatifs des innovations qui sont introduites par l'évolution des pratiques, qu'elles soient médicamenteuses, techniques ou organisationnelles. Bien sûr, celles-ci ont déjà été étudiées avant leur adoption et leur

autorisation de mise sur le marché, par exemple lors d'essais cliniques, mais sur des effectifs limités. Lors de leur diffusion et adoption à plus grande échelle, diverses questions peuvent se poser : de transposabilité de l'essai, des effets. Certaines innovations font aussi l'objet d'obligation de suivi après la mise sur le marché, mais pas toutes, et ces suivis ne sont pas aisés à mettre en œuvre, sans accès aux données de vie réelle fournies par les bases de données administratives. Ceci est particulièrement vrai pour les dispositifs médicaux dont les bénéfices sont rarement documentés par des essais cliniques, contrairement aux nouveaux médicaments.

2.5. Bénéfices pour l'amélioration de l'accès aux services de santé

L'analyse des données de santé peut permettre de documenter l'accès des patients aux interventions de santé : on peut calculer des taux et délais, et les croiser avec des variables géographiques, des données sociales et économiques, et ainsi identifier ce qui fonctionne ou non, et sur quels territoires doivent porter les efforts d'amélioration en priorité, même si le chemin peut être long entre l'identification d'un problème et l'implémentation des mesures correctives, particulièrement dans le champ du handicap et de la santé mentale par exemple.

2.6. Bénéfices pour l'amélioration des connaissances

Les données de santé contribuent aussi directement à l'amélioration des connaissances, étape indispensable pour améliorer les pratiques et concevoir des politiques publiques efficaces et protectrices. Pour beaucoup de maladies peu fréquentes, *a fortiori* rares, leur histoire naturelle est mal connue. Les registres ou cohortes spécialisés dans une maladie n'ont que très rarement un recrutement national, et le plus grand mal à ne pas perdre de vue les malades qu'ils incluent, empêchant la compréhension de l'évolution à long terme (17). Les données de santé administratives ont la capacité de combler les lacunes du savoir généré par les registres et cohortes.

2.7. Bénéfices pour le développement d'outils et de services améliorant l'état de santé de la population

Le système de santé utilise des outils de plus en plus performants. Les techniques d'investigation et de diagnostic, et les modalités de prise en charge et de traitement, se développent à partir de la recherche fondamentale et appliquée académique, et se transforment en produit à la disposition de tous, par des entreprises qui en assurent le développement, la production et la diffusion. La disponibilité des données de santé est un élément pivot de l'écosystème permettant l'innovation industrielle, les industriels pouvant ainsi mesurer la nature et la taille des besoins à couvrir, collaborer avec les professionnels de santé et les chercheurs académiques pendant la phase de recherche et développement, et évaluer leurs produits après la mise sur le marché.

Parmi les outils en développement pour améliorer la qualité des prises en charge, figurent tous les outils utilisant l'intelligence artificielle. Les algorithmes ne peuvent se développer que si des masses considérables de données sont disponibles. Les données utilisées doivent être celles indispensables à la finalité du traitement, pas plus. Les algorithmes ont à faire la preuve de leur utilité comme n'importe quelle innovation. Une plus grande disponibilité des données permettra le développement d'algorithmes de meilleure qualité car entraînés et testés sur des populations représentatives, ce qui n'est pas toujours le cas actuellement.

2.8. Bénéfices pour guider la politique de santé

Les données de santé sont cruciales pour élaborer les politiques de santé et les évaluer si elles sont effectives. Elles permettent d'évaluer l'efficacité des campagnes de prévention ou de dépistage, et l'observance des recommandations de bonnes pratiques émises par la Haute Autorité de Santé, l'Agence Nationale de L'Agence Nationale de Sécurité du Médicament et des produits de santé, ou les sociétés savantes. Elles permettent d'identifier les inégalités territoriales, sociales ou de classe d'âge pour envisager des mesures correctives. Elles sont clés dans la détection des tendances haussières de certaines pathologies, dont les épidémies. Toutes ces activités ne sont pas possibles sans l'exploitation des données de santé administratives. L'exploitation de données disponibles à tous les chercheurs pendant la pandémie de SARS-Cov 2 a fait la preuve de l'utilité d'une large ouverture des bases de données et montré que les données françaises étaient moins ouvertes que dans les pays anglo-saxons et d'Europe du Nord.

C'est dans cette optique qu'a été constitué le Comité stratégique aux données de santé auprès du ministre chargé de la santé. Ce comité apporte au ministre des éléments d'orientation et de décision relatifs à la mise en œuvre et au développement du système national des données de santé. Prévu par l'article R. 1461-10 du code de la santé publique, le comité stratégique est notamment chargé de proposer des orientations sur le développement du SNDS ; d'émettre un avis sur les évolutions législatives et réglementaires ; d'identifier des bases de données qui doivent être inscrites au catalogue et des catégories de données manquantes ou encore de réunir des groupes de travail thématiques et auditionner les responsables de bases de données pressentis. Sa présidence déléguée est assurée par la Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (Drees), sa vice-présidence par la Direction générale à la recherche et l'innovation (DGRI) et son secrétariat par le Health Data Hub.

Récemment, le Comité stratégique des données de santé a notamment mis sur pied plusieurs groupes de travail représentant de l'écosystème, visant à définir un socle commun de données pour les entrepôts de données hospitaliers, une gouvernance d'accès aux données de santé ou encore des modalités de financement pérenne des bases de données de santé et mécanismes de redevance.

2.9. Bénéfices pour la recherche participative

Les sciences et recherches participatives sont des formes de production de connaissances scientifiques auxquelles participent, aux côtés des chercheurs, des acteurs de la société civile, à titre individuel ou collectif, de façon active et délibérée. Elles sont un moyen d'impliquer les citoyens dans la recherche scientifique par l'association de l'expertise citoyenne et de l'expertise scientifique. C'est un outil concret pour répondre aux enjeux sociétaux et bénéficier d'un savoir expérientiel très large, qui complète harmonieusement le savoir expert des acteurs de la recherche. La disponibilité de larges jeux de données est une des conditions du développement de la recherche participative. Du côté des recherches, le regard et l'expérience des usagers sont complémentaires de ceux des experts. Du côté des usagers et citoyens, leur implication leur permet non seulement de dialoguer avec les experts, de monter en connaissance dans le champ de la recherche et des montages de projet, de la compréhension de l'utilisation des données de santé et du cadre réglementaire associé. L'objet est non seulement le déploiement de l'effort des structures publiques pour créer en cohérence avec leurs missions une gouvernance citoyenne adaptée et des espaces de dialogue, mais aussi de mettre en œuvre des projets de sensibilisation, d'information, de formation, voire des ateliers d'utilisation des données de santé.

3. Les effets délétères potentiels et leur risque de survenue

Après cette revue des bénéfices attendus de l'utilisation des données de santé disponibles, et qui seraient impossibles sans elles, il convient de regarder attentivement les risques générés ou potentiels qui sont résumés dans le tableau 2.

3.1. Risque de révélation de données personnelles

Les données de santé d'une personne constituent des informations très intimes qui ne doivent pas être divulguées au-delà du cercle très restreint des professionnels du soin choisis par la personne. Cet avis est consensuel et ne souffre pas d'exceptions. Il faut donc protéger le plus efficacement possible l'identité des personnes dont on souhaite utiliser les données. L'utilisation des données pour autre chose que le soin, pour la recherche donc, ne peut se faire que sur des données anonymes ou pseudonymes, les noms étant transformés en pseudonymes attribués aléatoirement sans possibilité de retrouver le nom à partir du pseudonyme. En effet, personne ne souhaite voir son dossier médical publié sur un site ouvert à tous.

L'enjeu est donc d'assurer une sécurité technique maximale contre le piratage des données des cabinets médicaux, des hôpitaux et cliniques et des bases de données administratives, c'est-à-dire de toutes les collections contenant des données nominatives. Et l'enjeu est de taille car le piratage des données nominatives est fréquent en raison de leur valeur monétaire élevée, soit pour l'obtention d'une rançon, soit par leur vente à d'autres acteurs économiques.

Les données utilisées pour la recherche ne sont, fort heureusement, jamais des données nominatives. Elles sont soit complètement anonymisées, soit pseudonymisées.

L'anonymisation est une technique qui permet de séparer les données nominatives et les données associées, de manière définitive. Aucun lien ne peut être reconstruit ultérieurement pour revenir à l'identité des personnes. C'est une façon radicale de rendre les données de santé utilisables par des chercheurs en protégeant parfaitement la vie privée. Cependant, ce système montre quelques failles. En effet, certaines données sont indirectement identifiantes par leur nature tout à fait unique ou très rare. Les données génomiques, particulièrement celles issues du séquençage, sont propres à chaque individu. Quelqu'un qui connaîtrait les caractéristiques du génome d'une personne, pourrait retrouver ses données dans toutes les bases génomiques. Cela demande une intention malveillante particulière et de gros efforts techniques mais c'est impossible à prévenir maintenant que beaucoup de personnes ont recours au séquençage de leur génome pour d'autres usages que la médecine et la science.

La pseudonymisation est une technique qui permet, en théorie, de protéger l'identité des personnes dont les données ont été collectées, en remplaçant le nom et le prénom par un pseudonyme attribué au hasard et ne contenant pas d'indication permettant de remonter à la personne. Elle a l'avantage de permettre la protection de l'identité des personnes sans compromettre un retour vers elles ou des données complémentaires les concernant si la recherche ou ses résultats le nécessitent. La correspondance entre le pseudonyme et les données nominatives est conservée dans une base de données à part. Les chercheurs n'ont accès, eux, qu'aux données liées aux pseudonymes et jamais à la table de correspondance. La pseudonymisation est par exemple une obligation réglementaire pour que le SNDS puisse être exploité par des chercheurs. Cependant, cette technique accuse quelques désavantages potentiels. Le plus courant est que le pseudonyme soit construit pour signifier quelque chose qui aide à la gestion des données au quotidien. Il arrive que des médecins mettent dedans les initiales du malade, l'année de recrutement, le titre du projet de recherche, et le numéro

d'inclusion du malade dans l'étude. Toutes ces indications facilitent beaucoup la réidentification des personnes par quelqu'un qui travaille dans l'institution par exemple, même si elles protègent suffisamment contre une identification par des personnes extérieures malveillantes. Ces mauvaises pratiques doivent être combattues.

Il existe d'autres failles permettant la réidentification par des personnes malveillantes totalement étrangères à l'institution, à condition qu'elles veuillent chercher une personne en particulier, que les données soient anonymisées ou pseudonymisées. Ceci peut se faire en croisant des bases de données dont les données sont publiquement disponibles. Si les fichiers de plusieurs sources contiennent l'année de naissance, le code postal de résidence, le sexe, le groupe sanguin, on peut en déduire qu'une personne en particulier a une pathologie précise parce qu'il figure dans un fichier des personnes atteintes de cancer par exemple. Ceci est arrivé en 1997 avec la réidentification du dossier médical d'un gouverneur aux USA par un journaliste. Les données génétiques humaines, les photos de visage, les enregistrements vocaux ou vidéos, l'imagerie médicale, peuvent permettre, en cas d'intention malveillante, une réidentification d'individus précis. Il n'y a pas, à ce jour, de cas connus de réidentification de données pseudonymisées à grande échelle. Les cas documentés sont le fait de personnes ayant cherché à démontrer que c'était théoriquement possible à partir de l'identification d'une personne particulière.

Des méthodes ont été développées pour diminuer la probabilité de réidentification. Elles reposent - par exemple - sur le principe de la confidentialité différentielle qui consiste à introduire du bruit de façon statistique pour altérer la précision des informations contenues. Elles ne font pas l'unanimité.

Aucune méthode ne peut prétendre empêcher totalement la réidentification d'une personne si des données en grand nombre sont disponibles et peuvent être croisées, mais une telle réidentification demande de gros efforts. Ainsi, le risque d'atteinte à la vie privée n'est pas complètement nul en cas d'usage malveillant des données à la fin de nuire, mais diminue drastiquement en cas d'usage encadré à des fins de recherche, si les règles de pseudonymisation sont bien respectées. Le risque est certes non nul mais il demeure faible et ne saurait justifier d'être considéré comme un obstacle à l'usage de données d'intérêt majeur pour la collectivité, considérant les bénéfices associés.

3.2. Risques d'utilisation non contrôlée

Les craintes exprimées portent sur le respect des droits des personnes concernées par l'utilisation de leurs données. La notion d'information préalable systématique est liée au fait que les données contiennent des informations sur la vie privée des personnes et qu'elle est un droit constitutionnel particulièrement protégé. Les données sont des objets publics et non privés, vis-à-vis desquels les détenteurs n'ont pas de droits mais des devoirs, devoirs de les protéger et de ne les utiliser qu'aux fins prévues lors de la constitution de la collection de données ou des traitements ultérieurs conformes au Règlement Général sur la Protection des Données (RGPD). Il faut focaliser la réflexion sur la collection de données et non sur les données individuelles (20).

Dans le soin, c'est un domaine qui a beaucoup évolué au fil du temps. En France, le principe général du consentement établi par le code de la santé publique est que les personnes doivent en principe consentir pour que leurs données soient recueillies dans le cadre d'une recherche sur leur corps. Ce principe s'est développé pour l'utilisation à des fins de recherche de prélèvements invasifs d'éléments du corps humain, la personne devant consentir pour autoriser un prélèvement qui ne lui servirait pas pour sa prise en charge médicale, mais qui ferait avancer la science. La participation à la recherche est vue comme

un don altruiste, d'éléments de son corps ou de temps passé à subir un test ou à remplir un questionnaire. Le consentement libre et éclairé doit rester la règle.

Mais dans l'utilisation de données personnelles collectées dans le cadre du soin, et qui seront utilisées à des fins de recherche, les personnes ne jouent aucun rôle supplémentaire qui puisse être assimilé à un don, les données ayant été collectées lors du soin pour les besoins du soin individuel et de la surveillance de la qualité du service rendu. C'est donc une situation différente qui ne doit pas, selon le RGPD et la loi informatique et libertés, nécessiter un consentement explicite et peut-être même pas une information systématique, puisque l'utilisation des données ne retire rien aux personnes et ne leur demande pas de contribution particulière. Pour autant il faut rassurer certains groupes comme les membres de minorités sexuelles ou les personnes soignées pour un trouble de santé mentale ou les femmes ayant eu recours à des actes d'orthogénie par exemple. Pour cela il faut que les institutions de soin soient irréprochables dans la protection des données qui leurs sont confiées, ce qui n'est manifestement pas encore totalement le cas. Une information sur la sécurité des données semble donc indispensable au moment du recueil des données, assortie d'une information sur leur utilisation secondaire.

Actuellement, la réglementation oblige par principe à la réalisation d'une information individuelle systématique, mais on pourrait légitimement mettre en place un système d'information générale sachant que les personnes disposent d'un droit de retrait du consentement sur requête, comme cela a été imaginé pour le don d'organes. Cela élargirait énormément le volume des données utilisables au bénéfice de progrès en santé publique, sans rien retirer aux bénéficiaires du système de santé.

Les données de recherche sont, elles, collectées avec l'aide active de la personne malade. Ces données sont généralement recueillies après consentement explicite après information détaillée sur les objectifs de la recherche et son protocole, car il s'agit d'un don pour la recherche (*opt-in*). Actuellement, en France, cette distinction n'est pas bien comprise ni appliquée, ce qui freine inutilement beaucoup d'études.

Il y a une autre crainte exprimée sur les usages non contrôlés, celui que les données servent à autre chose que les usages prévus dans l'intérêt général, en particulier par l'industrie qui pourraient avoir des finalités mercantiles, mais aussi par des groupes qui auraient des finalités partisanes ou de ciblage malveillant. Ce risque existe en théorie mais il est remarquablement contrôlé par la réglementation actuelle qui exige l'avis d'un comité indépendant sur la finalité des recherches, seules les recherches au service du bien commun sont autorisées. De plus, la réglementation européenne prévoit de lourdes amendes en cas de non-respect du règlement général sur la protection des données de l'Union Européenne.

3.3. Risque de vol de données et d'espionnage industriel ou étatique

La troisième classe d'effets délétères potentiels fait référence aux conséquences potentielles d'un vol de données de santé pour espionnage industriel ou étatique. Cette crainte est très présente dans le discours des opposants à l'ouverture des données de vie réelle pour la recherche, au prétexte que beaucoup d'entrepôts de données stockent leurs données dans un cloud géré par une entreprise américaine. En effet, le Cloud Act américain dispose que l'Etat américain peut accéder à des données situées n'importe où à l'étranger, en cas de nécessité lié à un danger pour la sécurité des USA. En fait, cette clause ne s'applique que dans certaines circonstances précises, par exemple des enquêtes judiciaires permettant à un juge d'enquêter pour retrouver des données de santé d'une personne accusée de terrorisme. Elle ne s'applique donc potentiellement qu'aux collections de données nominatives, ce qui n'est pas le cas des données conservées pour la recherche.

De plus, les entreprises américaines opérant en Europe doivent respecter le RGPD, et elles le font car elles risquent une amende de 4 % de leur chiffre d'affaires annuel en cas de non-respect du RGPD. De plus, ce précédent leur ferait perdre tous leurs clients européens.

Si beaucoup d'institutions publiques hébergent leurs données dans des clouds américains, c'est parce que les solutions technologiques américaines offrent des outils aux meilleures normes techniques et de sécurité. Ils sont les seuls à pouvoir répondre aux exigences des appels d'offres. Il n'existe donc pas d'alternative technologique satisfaisant aux normes de sécurité de haut niveau, ni en France, ni en Europe. Il apparaît plus important de privilégier la sécurité informatique, plutôt que la protection contre un hypothétique risque d'appropriation par les USA de données pseudonymisées, ce qui ne constituerait pas un dommage pour les citoyens ni individuellement, ni collectivement. Cependant, le développement d'un cloud européen est hautement souhaitable pour établir notre souveraineté dans le domaine.

En revanche, les données de soins primaires nominatives peuvent être volées par des hackeurs privés ou opérant pour des puissances étrangères. Il existe d'ores et déjà un marché économique pour ces données. Cette menace est une réalité pour les collections de données primaires, mais n'en est pas une pour les collections de données pour la recherche.

3.4. Risque de violation de la propriété intellectuelle et de préjudice financier par obstacle à la valorisation

Les professionnels impliqués dans la collecte, la gestion et l'analyse des données de santé, et les institutions qui les emploient revendiquent des droits de propriété intellectuelle sur ces données. Beaucoup s'en sentent même propriétaires, ce qui est un abus de langage. En effet, la donnée de santé elle-même ne fait pas l'objet d'un droit de propriété par son détenteur ni même le patient, mais il existe un droit de propriété intellectuelle sur les bases de données. Il existe également des protections liées au plagiat des œuvres. Cependant, les collecteurs de données font face à un certain nombre d'obligations : ils sont responsables du bon usage des données qu'ils collectent et de la production de valeur à partir de ces données. Pour être efficace, cette collecte est difficile et exige des investissements financiers ainsi qu'un haut niveau de professionnalisme, et les collecteurs peuvent en tirer une réticence à partager leurs données avec des tiers qui n'ont pas participé à l'effort de constitution. À cela s'ajoute l'illusion que leur collection a une valeur d'usage très importante, en raison de la médiatisation des profits réalisés par les grands opérateurs privés à partir de données nominatives, donc utilisables pour le ciblage commercial. Les données pour la recherche n'ont de valeur commerciale que si elles répondent à des standards de qualité internationaux et si elles sont collectées en grand nombre, permettant de mener des études puissantes.

La valorisation des collections de données pour la recherche est notoirement insuffisante, sans une ouverture des données à des tiers, académiques ou industriels, et sans chaînage des données avec d'autres collections existantes. Les tensions dans ce domaine peuvent être résolues en finançant correctement les équipes qui entretiennent des collections d'intérêt, ce qui n'est pas le cas actuellement, en travaillant des modalités de partage leur permettant d'obtenir un bénéfice scientifique et financier en cas d'ouverture des données à des tiers. Il existe donc des solutions concrètes pour répondre aux inquiétudes des acteurs, et aucun effet délétère potentiel avéré.

De plus, les grandes collections de données aux normes sémantiques et techniques ont une valeur d'usage indiscutable comme le montre l'expérience d'autres pays, en particulier Israël, Singapour, l'Estonie et les USA avec le réseau des Mayo clinics. Il n'y a pas

actuellement de modèle économique défini en France, où l'Etat est le financeur quasi exclusif. Il conviendrait donc de s'accorder sur la définition d'une politique publique de financement pérenne des bases et entrepôts permettant d'assurer leur collecte ainsi que les investissements initiaux qui ne pourront être couverts par des redevances, et de rassurer les acteurs. Ainsi financée, une collecte, une mise en qualité et une mise à disposition efficace permettrait aux données de santé d'alimenter l'économie de la connaissance sans dépendre d'acteurs privés. Ceci répondrait également aux attentes du monde industriel qui a besoin d'accéder à ces données dans le processus de R&D avec un enjeu de rapidité important.

3.5. Risque de surveillance normative des professionnels de santé

L'exploitation des bases de données de pratiques est vécue par certains acteurs comme une forme de contrôle de leur activité professionnelle inacceptable, car pouvant leur porter préjudice au travers de scores de performance rendus publics ou de sanctions professionnelles en cas d'écart notable par rapport à la moyenne. Ces types d'exploitation des données sont légitimes de la part des payeurs de ces services pour assurer l'usage optimum de l'argent public. La réponse à ces craintes des acteurs ainsi évalués demande une implication de ceux-ci et des institutions concernées dans l'interprétation des résultats, comme il a été fait lors de l'introduction des contrôles de qualité externe des laboratoires.

3.6. Risque de profilage des utilisateurs du système de soin

L'exploitation des données de l'Assurance Maladie ouvre la possibilité d'identifier des individus ou des groupes abusant de services ou porteurs de pathologies transmissibles et ne se soumettant pas aux obligations sanitaires en vigueur, ou ayant des profils atypiques considérés à risque pour certaines pathologies. Ce risque est réel et peut conduire à justifier des propositions politiques discriminantes. La protection repose sur le maintien d'une gouvernance démocratique de l'usage des données.

Après cette revue des bénéfices potentiels et des effets délétères possibles, il convient de décrire maintenant l'écosystème français et européen de gestion des données construit pour protéger les données de santé, dans ses dimensions techniques et réglementaires.

4. La gouvernance actuelle des données de santé

La collecte, la gestion et l'utilisation des données de soin, comme des données de recherche, sont très encadrées pour répondre aux souhaits et craintes des populations. La nature des réponses est technique, juridique et organisationnelle. Elles sont maintenant présentées.

4.1. Protection technique des données

La sécurité des collections de données de santé est un problème organisationnel et technologique majeur, sous la surveillance de l'Agence nationale de la sécurité des systèmes d'information (ANSSI) et du Haut fonctionnaire de défense et de sécurité du ministère de la Santé et de la Prévention. Les mesures de sécurité préconisées pour protéger l'accès aux données de santé sont les suivantes : les données doivent être chiffrées ; les droits d'opération doivent être segmentés entre les personnes de telle sorte qu'un éventuel acte de malveillance puisse être contenu ; la gestion des comptes et des permissions doit être sécurisée par une authentification reposant sur la combinaison de plusieurs facteurs d'authentification afin d'éviter toute usurpation d'identité ; les utilisateurs doivent disposer

d'un espace de travail sécurisé dédié uniquement à leur projet ; les briques techniques de sécurité doivent être indépendantes et utiliser des solutions de filtrage de flux, de détection de logiciels malveillants, de génération de clés de chiffrement ; tous les accès doivent être tracés ; les données peuvent en outre être hébergées sur le territoire de l'Union européenne par un hébergeur certifié "Hébergeur de données de santé".

Si tous ces critères sont respectés, les données sont en sécurité autant que faire se peut. Elles ne peuvent être accédées que par des actes criminels contre lesquels il n'y a pas de bouclier absolu.

Il existe donc un risque théorique de vol de données en cas de faille technique de sécurité. C'est un problème pour les hébergeurs de données de soin qui sont nominatives et ont donc de la valeur marchande, mais ce danger est pondéré dans le cadre de la recherche en santé par les mesures de pseudonymisation et d'anonymisation appliquées aux données. De plus, les collections de données pour la recherche ne sont généralement que des copies de bases de données qui sont hébergées par leur gestionnaire. Il ne peut donc pas y avoir de risque de pertes de données.

Le principal danger est lié au haut niveau de sécurité nécessaire à mettre en œuvre pour contrer le piratage informatique. Nombre d'institutions ne disposent ni du savoir-faire ni des budgets pour sécuriser leur système informatique, particulièrement les établissements de soin.

4.2. Protection juridique des données personnelles

La protection des données personnelles est réglementée par le Règlement général des données personnelles (RGPD) et certaines exigences sont précisées dans la Loi informatique et libertés (LIL) s'agissant des données de santé. Les projets de recherche mobilisant des données personnelles de santé sont réalisés sous le contrôle de la CNIL : dans certains cas, ils font l'objet d'une demande explicite d'autorisation qui lui est adressée, dans d'autres ils doivent se mettre en conformité avec des référentiels simplifiés[2] que la CNIL a élaboré. Cela signifie qu'ils ne sont pas obligés de recourir à une autorisation formelle à condition qu'ils vérifient toutes les conditions listées dans ces référentiels. La CNIL est susceptible de le vérifier.

Par ailleurs, les textes prévoient également des exigences de sécurité à respecter pour l'hébergement, c'est par exemple le cas du référentiel de sécurité du SNDS, des exigences de transparence sur les projets et leur résultat et l'obligation d'information et d'accompagnement des citoyens qui souhaiteraient exercer leur droit de retrait ou de modification.

4.3. Évaluation des finalités de la recherche

Il y a un consensus pour estimer que l'utilisation des données de santé pour un projet de recherche n'est légitime que si la finalité de la recherche est légitime et éthique, et la méthodologie appropriée, mais aussi que le projet a un caractère d'intérêt public et que les porteurs de projet en ont véritablement la légitimité.

Cette vérification est réalisée par le Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES), remplaçant le comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé

[2] Les "référentiels simplifiés" ou "méthodologies de référence" font référence à une simplification des démarches d'accès aux données de santé. Ils facilitent les traitements de données personnelles de santé puisqu'une autorisation expresse de la CNIL ne serait pas requise pour les traitements qui s'y conformeraient. cf. Cnil : https://www.cnil.fr/fr/les-referentiels-et-methodologies-de-reference-sante

(CEREES). Il a été créé à la parution du décret d'application de la loi informatique et libertés, le 15 mai 2020, et de l'arrêté de nomination de ses membres en date du 9 juin 2020. Il est institué auprès du ministre chargé de la Recherche et du ministre chargé de la Santé. Pour formuler son avis, il vérifie notamment si le projet est utile socialement ; si le projet est sérieux et crédible ; que le porteur de projet ne demande pas l'accès à plus de données que nécessaires pour le projet ; qu'il y a un retour à la société civile pour avoir partagé ses données. À cette fin le comité est composé d'une part d'une vingtaine de membres nommés par arrêté, comprenant notamment deux représentants des usagers de santé et d'autre part d'un réseau d'une quarantaine d'experts désignés par le Président sur proposition éventuelle des membres. Cette composition permet de disposer des expertises nécessaires pour examiner les dossiers soumis, c'est-à-dire non seulement méthodologiques, mais également cliniques.

Le CESREES donne un avis, avant examen par La CNIL qui reste la seule institution habilitée à autoriser des porteurs de projets à traiter les données. La CNIL vérifie que le projet respecte toutes les obligations de sécurité exigées par le Règlement général sur la protection des données.

4.4. Cadrage juridique des partenariats

Lorsque les données des acteurs sont partagées, il est possible d'encadrer les modalités de partage par le biais de contrats entre notamment le producteur de données et l'utilisateur, ou encore entre le producteur et la plateforme de partage ou entre l'utilisateur et la plateforme de partage.

Ces contrats rappellent les obligations en matière de respect du RGPD, dans le cadre du transfert de données et de la recette ou de la correction des données le cas échéant. Les contrats couvrent aussi les mesures prises pour mettre en visibilité la base de données (mise en ligne dans un catalogue de métadonnées, attribution de DOI [*Digital Object Identifier*, un élément numérique persistant permettant d'identifier les données] par exemple) et précisent les modalités de valorisation scientifique des parties et notamment les conditions de publication. Ils peuvent aussi inclure les modalités d'enrichissement des données par d'autres, d'ouverture de certains résultats et les licences retenues, les modalités tarifaires ou encore les financements associés à la réalisation des travaux requis pour mettre en œuvre le partage le cas échéant. Les contrats liant les utilisateurs impliquent aussi des conditions générales d'utilisation des plateformes et des exigences de sensibilisation. Ces contrats sont établis entre les parties qui doivent arriver à un consensus avant que le projet ne commence. Les institutions et les chercheurs restent donc maîtres de l'utilisation de leurs données par des tiers. Le problème principal de ce secteur est le manque de confiance réciproque entre institutions publiques qui est source de retards importants, voire même de blocage, dans la signature des contrats (21).

4.5. Implication des citoyens dans les choix

Plusieurs mesures sont prises par la loi pour impliquer le citoyen dans la gouvernance d'accès aux données de santé. À titre d'exemple, il est prévu par la loi que le Health Data Hub, opérateur national réunissant 56 parties prenantes de l'écosystème des données de santé, soit vice-présidé par le président de France Assos Santé et que deux représentants des usagers des données de santé siègent au CESREES ainsi qu'au Comité stratégique des données de santé, organe créé par la loi pour fixer les grandes orientations du SNDS.

4.6. Politique générale de l'utilisation des données de santé en France

Consciente des bénéfices de la réutilisation des données de santé, la puissance publique a ouvert la voie de leur ouverture en 2016 avec la loi Modernisation de notre système de santé (LMSS).

Cette loi marqué une étape majeure. En créant le SNDS, elle a unifié la gouvernance de plusieurs grandes bases de données administratives qui étaient antérieurement gérées par des organismes différents, avec des règles d'accès aux données spécifiques à chacune d'entre elles. La gouvernance stratégique de ce patrimoine de données est désormais assurée par l'État.

Elle a explicité la doctrine et les principes concernant l'accès aux données de ce SNDS : tous les acteurs, publics ou privés, peuvent y accéder, dans des conditions assurant leur sécurité et la protection de la vie privée, dès lors que les finalités de leur utilisation ne sont pas contraires à l'intérêt public. Des règles et modalités d'accès ont été définies, et un guichet unique a été créé pour faciliter ces accès, l'Institut national des données de santé (INDS).

La loi pour une Organisation et transformation du système de santé (OTSS) du 24 juillet 2019 a renforcé cette ambition de développement des usages au service de l'innovation et de la recherche. En étendant le Système national des données de santé à toutes les données de santé associées à un financement public, elle pose le principe que celles-ci ont vocation à être plus largement utilisées. Elle a remplacé l'INDS par le Health Data Hub qui, au-delà du rôle de guichet unique, assure des missions beaucoup plus larges :

- la mise à disposition d'une plateforme sécurisée à l'état de l'art pour stocker et traiter les données ;
- l'élaboration et l'enrichissement progressif d'un catalogue de données documenté pour mettre à disposition de la communauté scientifique, au-delà du SNDS « historique », des bases de données jugées prioritaires pour faire avancer les connaissances en santé, cohortes, registres, données hospitalières... Le Health Data Hub est co-responsable de traitement de la base principale du SNDS avec l'Assurance maladie et responsable de traitement du catalogue ;
- des services et outils pour les utilisateurs, l'animation de l'écosystème pour accélérer l'innovation en favorisant le partage d'expériences et de connaissances.
- l'animation des travaux du Comité stratégique des données de santé qui a pour missions de proposer des orientations sur le développement du SNDS, émettre un avis sur les évolutions législatives et réglementaires, identifier des bases de données qui doivent être inscrites au catalogue et des catégories de données manquantes, réunir des groupes de travail thématiques et auditionner les responsables de bases de données pressentis.

4.7. Politique générale de l'utilisation des données de santé en Europe

Pour faciliter l'accès aux différents types de données disponibles dans les États membres, la Commission européenne a fait du futur espace européen des données de santé (EHDS) l'une des priorités de sa politique de santé (22). Cette ambition se traduit notamment par la proposition d'un règlement européen créant l'espace européen des données de santé, couvrant l'utilisation primaire et secondaire des données de santé. Le projet de règlement est actuellement en négociation au Parlement européen et au Conseil de l'Union européenne en vue d'une adoption prévue en 2024. Pour préparer l'implémentation du texte, plusieurs instruments de préfiguration du futur espace européen des données de santé ont été mis en

place depuis 2019. C'est le cas de TEHDaS (*Towards a European Health Data Space*) (23), un programme de réflexion réunissant plus de 26 États membres dans lequel le Health Data Hub a coordonné cinq partenaires français. Pilotée par le HDH, la contribution française a été particulièrement active sur l'engagement citoyen et la gouvernance. En juillet 2022, la France a pris la tête du consortium lauréat d'un appel à candidatures de la Commission européenne pour construire une version test du futur espace européen des données de santé. Les travaux ont été lancés en octobre 2022 et devront durer deux ans. Il s'agit de proposer un parcours utilisateur de bout en bout pour un chercheur souhaitant utiliser des données de santé de plusieurs pays européens. Des services proposés dans ce parcours utilisateurs incluent un catalogue de métadonnées européen, alimenté par des catalogues nationaux, et un formulaire de demande d'accès unique, pour éviter qu'un chercheur doive soumettre de multiples demandes d'accès en parallèle.

Le HDH français est vu comme un exemple à suivre pour les autres pays européens, et la Commission européenne s'est inspirée du modèle français en proposant la création dans chaque pays d'organismes responsables de l'accès aux données de santé. Récemment, la Commission européenne a approuvé le projet porté par le Health Data Hub et ses partenaires pour préparer la mise en œuvre nationale de l'Espace européen des données de santé ("*French HealthData@EU*"). Le consortium français percevra un financement afin de préparer la mise en oeuvre de l'EHDS au niveau national avec notamment : 1) un renforcement des services du HDH ; 2) des fonds qui impulsent une démarche coordonnée pour la mise en qualité et la standardisation des données en vue de leur future réutilisation et 3) le remplissage du catalogue de métadonnées national par les partenaires détenteurs de données, respectant le standard européen dans la matière (Health DCAT-AP).

4.8. Le processus d'accès aux données et ses freins

Malgré les évolutions prises ces dernières années, des freins subsistent au partage de données.

Les démarches d'accès sont nombreuses et complexes, il n'est pas forcément facile pour les acteurs de savoir quelles démarches réaliser entre les demandes d'autorisation, et si leur projet correspond aux référentiels simplifiés ou à des méthodologies de référence. Par ailleurs, ces dernières évoluent régulièrement sans que la stratégie d'élaboration de ces référentiels soit très transparente.

Le partage des rôles entre les ministères et la CNIL sur l'élaboration de la gouvernance nationale d'accès aux données est d'une manière générale flou. L'usage du NIR (répertoire national d'identification des personnes physiques, c'est à dire le numéro de sécurité sociale attribué à chaque personne à sa naissance, sur la base d'éléments d'état civil transmis par les mairies à l'INSEE), les règles d'anonymisation, les démarches de mise en conformité des bases de données ou encore la bonne manière de mettre en œuvre les exigences en matière d'information et d'exercice des droits, posent encore des problèmes. Dans certains cas, le franchissement des différents jalons réglementaires, contractuels ou techniques se traduit par des délais d'accès longs et dissuasifs qui peuvent se chiffrer en années. Ceci conduit à des situations absurdes où les données deviennent disponibles alors que le financement du projet de recherche a pris fin et que les chercheurs sont partis travailler ailleurs. Ces situations représentent de vraies pertes de chance car des projets utiles et financés ne se font pas pour des délais administratifs sans justification. Aucune étude quantifiant l'ampleur de ces problèmes et leur retentissement sur la recherche n'est disponible. Cependant la frustration de la communauté des chercheurs s'exprime souvent. La récente publication d'un livre blanc du Conseil Scientifique du CNRS, sur les obstacles administratifs que les chercheurs rencontrent au quotidien, révèle une profonde souffrance (21) qui appelle des

réponses. Un audit des process paraît indispensable, car le système de protection des données s'est construit progressivement, à une époque où le concept de science ouverte n'était pas adopté comme une valeur. La multiplication des acteurs est en soi un problème, source de délais importants non justifiés. Il faut donc repenser le système d'autorisation avec une seule instance délivrant les autorisations de traitement, comme le prévoit la directive européenne en préparation sur l'espace européen des données de santé.

Les exigences en matière de sécurité ou, plus récemment, de souveraineté sont très élevées et déconnectées du terrain, que ce soient les moyens que les acteurs sont en capacité de mobiliser pour le partage et l'accès aux données de santé ou le niveau des services disponibles sur le marché privé. La CNIL exige régulièrement que les données de recherche ne soient pas stockées dans un cloud géré par un acteur américain alors qu'il n'existe aucun autre cloud français ni européen remplissant le cahier des charges sécuritaires exigé.

De plus, l'absence d'autorisation par la CNIL d'utiliser un identifiant unique rend les croisements parfois complexes et moins efficients, et ajoute plusieurs mois de travail pour réaliser des appariements statistiques. Beaucoup d'autres pays européens réalisent des appariements à partir d'un identifiant unique en routine dans le respect du RGPD. Ceci met les équipes de recherche française en situation d'infériorité vis-à-vis de ses compétiteurs, retarde beaucoup les projets de recherche et rend les études moins puissantes puisque les appariements statistiques ont des rendements de l'ordre de 80%, un taux qui pourrait atteindre près de 100% avec la mobilisation du NIR. On peine à trouver la justification à une telle interdiction, car cela n'augmente en rien la protection des données.

La mise en œuvre de grandes bases de données de santé nécessite d'importantes dépenses pour les infrastructures techniques, les développements informatiques et le développement de logiciels métier par exemple. Les institutions doivent aussi documenter, standardiser, pseudonymiser, apparier, mettre en qualité les données pour qu'elles soient utilisables par des tiers. Elles doivent aussi mettre en œuvre les mesures d'information des personnes ou l'exercice des droits prévu par le RGPD. Il faut aussi qu'elles aient toute l'organisation nécessaire pour contractualiser et mettre à disposition les données à des tiers dans le respect des droits et sous haut niveau de sécurité. Ces coûts sont souvent sous-estimés par les administrations en charge et les décisions stratégiques difficiles à prendre en raison de la frilosité générée par la question du partage des données de santé pour les décideurs politiques. Pour que ceci se fasse, il faut un fort soutien aux producteurs de données de santé d'intérêt et une aide méthodologique, technique et réglementaire actuellement fournie par le Health Data Hub.

5. Les services offerts par le HDH pour répondre à ces tensions éthiques

Les missions et projets du Health Data Hub ont été développés dans le respect des valeurs de la science ouverte, de l'intégrité scientifique, de la valorisation sociale, de la protection des données personnelles, de la progression des connaissances au service de tous et de l'optimisation énergétique des ressources écologiques, financières et humaines.

Le principal besoin exprimé par les acteurs en France est celui de la lisibilité des bases de données disponibles, de leur contenu et des règles d'accès aux données. Ils souhaitent également pouvoir y accéder sous des délais courts au risque de leur préférer des données disponibles plus simplement ou plus rapidement dans d'autres pays.

Dès 2016, la France a mis en place un acteur national unique pour accompagner les porteurs de projets en ce sens. Bien que renforcé par la loi de 2019, il n'a jamais eu vocation à être le seul acteur capable de mettre effectivement à disposition les données. Plus que jamais, les acteurs expriment leur besoin d'avoir un interlocuteur capable de les assister dans le processus réglementaire : de l'identification de la procédure d'accès applicable à la préparation du formulaire de demande en passant par la liaison avec le comité éthique et scientifique.

En France, le HDH assure le secrétariat du comité éthique et scientifique national et communique les dossiers de demande à la CNIL. Les détenteurs de données peuvent toujours disposer d'un comité éthique et scientifique à un niveau plus local, mais l'avis de ce comité fera partie du dossier soumis à la gouvernance nationale si une autorisation de la CNIL est requise. La CNIL est aujourd'hui le seul acteur mandaté pour autoriser le traitement des données, garantir le traitement équitable de tous les dossiers et la transparence du processus et des avis rendus.

Le HDH anime la communauté des utilisateurs de données, identifie ses besoins en termes d'accompagnement, produit de la documentation pédagogique et promeut l'approche open source. Il met aussi en relation les utilisateurs de données avec les acteurs institutionnels compétents, en fonction des difficultés qu'ils rencontrent, et peut parler en leur nom aux acteurs institutionnels en cas de difficultés rencontrées et résoudre concrètement des blocages dans la mesure du possible.

Le HDH est surtout l'opérateur national pour la mise en œuvre des grandes orientations stratégiques identifiées et peut, en particulier, mener les travaux priorisés par le comité stratégique. Le HDH facilite les travaux au niveau national sur l'harmonisation des tarifs d'accès, l'identification de jeux de données minimaux, la mise en place de procédures d'accès simplifiées, la mise en place d'un système d'information sur les données de santé, etc.

Le HDH soutient et guide les acteurs dans la mise en œuvre de mesures liées à l'ouverture des données de santé : par exemple, en France, une politique est actuellement mise en œuvre pour soutenir la conception et le déploiement d'entrepôts de données de santé hospitaliers. Le Health Data Hub est partenaire de cette initiative depuis la conception jusqu'à l'accompagnement des lauréats.

Le HDH centralise l'information pour maximiser la transparence vis-à-vis de la société civile (projet de formulaire numérique national permettant aux citoyens de faire des demandes d'exercice de leurs droits de manière simplifiée et traitée - en cours de construction). Il contribue à la sensibilisation et à la formation (initiale et continue) des professionnels de santé à l'utilisation secondaire des données, afin qu'ils puissent favoriser la constitution et le partage de bases de données de santé d'intérêt pour la recherche et l'innovation.

Il met à disposition les données des différents détenteurs de données en agissant en tant que tiers de confiance. La quantité de données transmises par son intermédiaire peut nécessiter des investissements importants en termes d'infrastructure technologique, qui n'auraient pas de sens au niveau local.

Le HDH permet de conserver les données d'intérêt national, afin de tirer parti des efforts déployés pour améliorer la qualité, le couplage et la disponibilité des données et réduire ainsi la charge que représente la fourniture de données pour les fournisseurs de données.

Il a, par ailleurs, été créé au sein du HDH une direction citoyenne pour élaborer un programme de travail favorisant la mise en œuvre de ses missions légales de transparence et d'accompagnement des citoyens. Ces actions concernent l'écoute des citoyens

(e-consultation, conférence de consensus), leur intégration au sein de la gouvernance et autour des réflexions (groupes de dialogue) ; l'information et la formation (une formation citoyenne a été élaborée avec FAS, la CNIL, la DREES et la CNAM) ; l'accompagnement dans l'exercice des droits avec notamment un formulaire national pour gérer les oppositions.

Conclusion

L'usage des collections de données de santé pour la recherche scientifique, médicale et en santé publique est pleinement justifié tant est grand le nombre des usages au service du bien collectif qui pourrait être ainsi effectués et qui ne le sont pas dans le cadre actuel.

Les risques potentiels pour les citoyens et la collectivité, sont très minimes si les mesures de protection qui ont été adoptées sont efficacement mises en œuvre, et aucun des risques n'est de nature à faire courir un danger à fort impact (de santé), ni individuel ni collectif. La balance bénéfices/risques penche donc clairement du côté des bénéfices et plaide pour une action corrective rapide des retards et blocages, injustifiés si on considère la balance bénéfice/risque et non exclusivement les risques théoriques. À titre d'exemple, on peut citer l'impossibilité pour les centres experts français des maladies osseuses constitutionnelles, de contribuer au registre international de la fibrodysplasie ossifiante progressive, une maladie très sévère et ultrarare qui conduit à l'ossification des muscles, pétrifiant au sens propre les malades, au prétexte que les données de ce registre sont stockées aux USA, car le registre a été créé par une association de malades américaine. Un stockage aux USA est considéré en soi par nos agences régulatrices comme un risque majeur, sans que la nature de ce risque soit imaginable, et alors que sa probabilité est nulle, car jamais observée jusqu'à ce jour après des dizaines d'années de collaboration scientifique transatlantique. Cette interdiction ralentit les connaissances sur l'histoire naturelle de la maladie, alors qu'elles sont indispensables pour évaluer des traitements innovants en développement.

Les finalités de l'usage des données en recherche sont très largement partagées dans nos sociétés, et alignées sur des choix déjà inscrits dans notre cadre juridique et législatif. Faciliter l'accès aux données pour la recherche est essentiel pour respecter les valeurs de la science ouverte inscrite dans la loi, pour répondre aux attentes des citoyens de bénéficier de soins appropriés car correctement évalués, pour faire progresser les pratiques par une meilleure connaissance des déterminants de la santé globale, pour assurer l'équité des territoires et des groupes sociaux, pour fournir les connaissances nécessaires à une bonne prise de décision à tous les niveaux, et à évaluer les politiques publiques.

Ces objectifs ne seront atteints que si les freins à la mise en œuvre de l'ouverture des collections de données sont levés et si la culture du bien commun que représentent les données est diffusée dans toutes les couches de la société. Les grandes collections de données de santé utilisées actuellement par les chercheurs (24) sont hébergées dans les pays tels que le Royaume-Uni ou les USA, et les institutions responsables de leur diffusion n'ont jamais rapporté d'effets néfastes attribuables au libre accès, contredisant les alertes lancées par les détenteurs de grandes collections de données en France.

La France peut s'enorgueillir d'avoir une politique aboutie de mise à disposition des données de santé pour la recherche, d'avoir construit un entrepôt national de données de santé et créé le HDH. Les bases administratives françaises sont uniques au monde en termes d'exhaustivité et de richesse d'information. Cependant des freins multiples ne permettent pas de mener à bien les multiples recherches nécessaires à un bon pilotage de la politique de santé, à l'amélioration du système de soin et à la recherche de solutions innovantes. On peut citer l'évaluation de chaque projet de recherche individuellement pour ses aspects de

sécurité technique et son respect des règles réglementaires, alors que ceux-ci dépendent de l'institution dans laquelle les chercheurs travaillent. Le bon sens voudrait que l'évaluation de ces services de gestion des données soit réalisée au niveau des établissements et non des projets. Cela simplifierait énormément les procédures, raccourcirait de plusieurs mois l'obtention des autorisations et responsabiliserait les instituts de recherche dans leur politique de gestion des données. C'est une revendication actuelle des Instituts Hospitalo-Universitaire (25). Beaucoup d'études ne se font pas ou avec grand retard du fait de difficultés concrètes d'accès aux données du SNDS. Cela ne peut durer car une solution est identifiée. Elle réside dans la capacité du HDH à héberger une copie du SNDS, dont la mise en oeuvre se heurte aujourd'hui au débat relatif à l'hébergement des données sur un serveur en France utilisant les technologies de Microsoft, une entreprise américaine, alors que le HDH devrait être considéré comme un réducteur de risques.

La priorité doit être donnée à la résolution des causes des retards importants rencontrés à l'obtention des autorisations de traitement, à la signature d'accords de partenariats entre producteurs de données, et à la mise à disposition des données de l'assurance maladie. Il n'est pas éthique d'empêcher des recherches utiles de se faire, par des exigences de sécurité allant bien au-delà de celles prévues par le RGPD, parfois impossibles à satisfaire et donc bloquantes (26).

Il faut donc sortir de la logique juridique pure qui ne prend en compte que les risques théoriques potentiels, sans considérer les bénéfices attendus de l'utilisation des données de santé pour la recherche. Cela nécessite de travailler à une grille d'évaluation des risques et des bénéfices pour atteindre un consensus sur le poids respectif de chacun des items. Cela est communément fait dans les processus de prise de décision pour les autorisations de mise sur le marché des produits de santé ou d'organisation de dépistages en population, par exemple. Les méthodes sont transposables.

Un audit de la situation actuelle serait bienvenu, pour déboucher sur un ajustement de notre système réglementaire qui doit protéger contre les risques impactant, sans entraver au prétexte de protéger de risques à effets négligeables et survenant exceptionnellement. Cela passe par la prise en compte de la balance bénéfice risque pour la délivrance d'autorisation ou de conformité, et non simplement des risques, comme actuellement par la CNIL. La seule instance qui prend en compte la balance bénéfices/risques est le CESREES. Afin de mieux appréhender les risques et les bénéfices des recherches secondaires sur données de santé pour la santé publique, nous proposons la mise en œuvre d'un outil d'évaluation du bénéfice/risque objectif, où les risques sont mesurés, pondérés par la probabilité d'occurrence du risque, objectivés par l'impact sur la société.

Trois autres chantiers doivent être ouverts, celui de la distinction du consentement des personnes entre données de soin et données de recherche, celui de l'usage du numéro d'identification national pour le chaînage des données et celui de la simplification du système des autorisations. Le rapport de la Commission des Affaires Sociales du Sénat va dans le même sens (27). Il recommande de faciliter les appariements par les chercheurs des différentes bases de données existantes en assouplissant le décret relatif aux professionnels autorisés à utiliser l'identifiant national et de clarifier les modalités de recueil du consentement des patients. Le rapport recommande aussi une clarification des missions de la CNIL et du HDH dans le domaine des autorisations. La levée de ces points de blocage changerait radicalement la situation et permettrait enfin de mener les études que le pays et les citoyens attendent légitimement.

Le rapport de l'action conjointe européenne *Towards the European Health Data Space* (TEHDAS) (23) note que "Malgré les efforts réalisés dans certains pays pour faciliter l'utilisation secondaire des données de santé, des obstacles persistent. En ce qui concerne

les projets transfrontaliers, les processus d'autorisation nationaux non harmonisés et largement non coordonnés restent un frein important." Il faut donc clarifier, rationnaliser et simplifier les étapes d'accès aux données de santé dans tous les États membres de l'Union européenne, dont en premier lieu la France qui a des exigences réglementaires supérieures aux exigences européennes. Pour y parvenir, les travaux de TEHDAS soulignent "l'importance d'une répartition claire des rôles et responsabilités incombant aux différents acteurs au niveau national : détenteurs de données, organismes responsables de l'accès aux données, et coordinateurs d'organismes d'accès aux données".

En outre, l'initiative européenne a fait valoir "le soutien des citoyens à la réutilisation des données de santé", confirmé par une grande consultation menée au Royaume-Uni, en Belgique et en France, avec notamment l'appui de France Assos Santé et du HDH qui a permis de recueillir plus de 6.000 contributions." (23) Les citoyens attendent que les données de santé servent à améliorer notre santé publique et notre système de soin. À nous de trouver les solutions pour que cela soit effectif. Nous avons besoin d'un cadre cohérent, fiable et efficace pour l'utilisation secondaire des données de santé, et c'est possible. Il faut encadrer sans entraver (28) et ne pas protéger quelques individus au détriment de progrès pour des millions de personnes.

Annexes

Tableau 1 : bénéfices à l'utilisation des données de santé à des fins de recherche

nature des bénéfices	bénéfices	impacts
bénéfices généraux	masse critique de données	décisif pour les évènements rares
		amélioration de la puissance des études
	représentativité de la population générale	amélioration de la pertinence des résultats
	mutualisation des compétences	professionnalisation des pratiques
	mutualisation des ressources techniques	amélioration de la sécurité des données
		économies d'échelle
		facilitation de la participation des petites institutions
surveillance épidémiologique	surveillance exhaustive sur le territoire	détection des tendances plus rapides
	amélioration de la quantité et de la qualité des données	meilleure mobilisation des registres
	diminution des coûts d'acquisition des données	identification des inégalités
amélioration des pratiques	mesure de l'efficacité des soins	meilleure allocation des ressources
	détection des inégalités territoriales	meilleur pilotage des politiques publiques afin de corriger les inégalités de soins
	pertinence des prises en charge	amélioration des pratiques professionnelles
évaluation des innovations	mesurer les effets des innovations en vie réelle	meilleure allocation des ressources
	justifier leur prise en charge ou non	éviter des scandales sanitaires
	détecter tôt les effets délétères	support à la décision publique
amélioration de l'accès aux services de santé	détecter les défaut d'accès et les surconsommations	améliorations pour un service équitable, territorial et sociologique
	détecter les inégalités territoriales	
amélioration des connaissances	meilleure compréhension de l'histoire naturelle des maladies	meilleure R&D
	identification des déterminants géographiques, économiques, sociaux, et environnementaux	innovations bien ciblées
		support à la décision publique

outils et services améliorant la santé	outils plus performants pour le diagnostic et la prise en charge	amélioration des prises en charge
	algorithmes d'aide à la décision, à la gestion et au suivi	amélioration des pratiques professionnelles
guidage pour l'élaboration des politiques	outils pour évaluer les politiques en place	meilleur pilotage des politiques publiques avec un meilleur ratio coût/efficacité
	suivi de l'effet des recommandations	réduction des inégalités territoriales
	mesure des inégalités territoriales	adaptation agile aux crises
recherche participative	implications des citoyens dans la recherche	meilleure adhésion aux politiques publiques
	prise en compte des savoirs des malades et des aidants	meilleurs services car plus adaptés aux attentes
	espaces de dialogue démocratiques	confiance dans les services

Tableau 2 : risques à l'utilisation des données de santé à des fins de recherche

nature des risques	conséquences potentielles	hauteur des risques
révélation de données personnelles	divulgation d'éléments de la vie privée	très bas pour les données pseudonymisées
		nul pour les données anonymisées
	réidentification malveillante	très bas pour les données pseudonymisées
		a fortiori nul pour les données anonymisées
utilisation non contrôlée	usage hors information des personnes si mauvaises pratiques	très bas car régulation forte en place
vol de données et espionnage	très peu de conséquence pour les données pseudonymisées ou anonymisées	extrêmement bas puisque jamais observé pour des données de recherche
violation de la propriété intellectuelle	perte de l'avantage d'antériorité	très bas car régulation forte en place
surveillance normative des professionnels de santé	sanction des professionnels qui s'écartent des pratiques standards	bas, car les écarts seront traités par la négociation
	avantages pour les payeurs qui veillent à la bonne allocation des ressources	
profilage des utilisateurs de soins	stigmatisation de certaines minorités	nul en démocratie mais élevé en démocratie illibérale

Références

1- Stratégie de transformation du système de santé, Gouvernement, août 2019. lien : https://www.gouvernement.fr/action/strategie-de-transformation-du-systeme-de-sante
2- Plan national pour la science ouverte, Ministère de l'enseignement supérieur et de la recherche, juillet 2018. lien : https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-tous-49241
3- "Big data en santé : des défis techniques et éthiques à relever", Inserm, juin 2022. lien : https://www.inserm.fr/dossier/big-data-en-sante/
4- "Recommendations on how to engage citizens in the European Health Data Space", Tehdas, 2023. lien : https://tehdas.eu/app/uploads/2023/03/tehdas-recommendations-on-how-to-engage-citizens-in-the-european-health-data-space.pdf
5- Health Data Hub, lien : https://www.health-data-hub.fr
6- Marcel Goldberg , Marie Zins, "Le Health Data Hub - Pourquoi ? Comment ?" Med Sci (Paris), mars 2021 pp.271-276. lien : doi: 10.1051/medsci/2021016.
7- Pierre Lombrail, Israël Nisand, Christine Dosquet, Frédérique Lesaulnier, Catherine Bourgain, et al, "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 "", Comité d'éthique de l'Inserm, janvier 2022. lien : inserm-03533863
8- "Données massives et santé : une nouvelle approche des enjeux éthiques", Comité Consultatif National d'Ethique, Avis 130 du CCNE, mai 2019. lien : https://www.ccne-ethique.fr/sites/default/files/2021-02/avis_130.pdf
9- "Plateformes de données de santé : enjeux d'éthique". Comité Consultatif National d'Ethique, Avis commun du CCNE et du CNPEN, Avis 143 du CCNE, Avis 5 du CNPEN, février 2023. lien : https://www.ccne-ethique.fr/sites/default/files/2023-05/CCNE-CNPEN_GT-PDS_avis_final27032023.pdf
10- "Qu'est-ce ce qu'une donnée de santé", Cnil, 2023. lien : https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante (consulté le 2 octobre 2023)
11- "Entrepôts de données de santé hospitaliers en France : Quel potentiel pour la Haute Autorité de santé ?", Haute autorité de santé, octobre 2022. lien : https://www.has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france
12- "P4DP, un consortium pour créer le premier entrepôt de données de santé pour la médecine générale", Health Data Hub, 16 mars 2023. lien : https://www.health-data-hub.fr/actualites/p4dp-un-consortium-pour-creer-le-premier-entrepot-de-donnees-de-sante-pour-la-medecine (consulté le 2 octobre 2023)
13- "France Cohortes : comment pérenniser un outil de recherche exceptionnel", Inserm, septembre 2020. lien : https://www.inserm.fr/actualite/france-cohortes-comment-perenniser-outil-recherche-exceptionnel/ (consulté le 2 octobre 2023).
14- "Registres et données de santé : utilité et perspectives en santé publique", Haut Conseil de la Santé Publique, septembre 2021. lien : https://www.hcsp.fr/Explore.cgi/avisrapportsdomaine?clefr=1126
15- Nathalie Blanpain, "L'espérance de vie par niveau de vie : Méthode et principaux résultats", document de travail n°F1801, Insee, 2018. lien : https://www.insee.fr/fr/statistiques/3322051

16- "Inégalités environnementales et sociales se superposent-elles ?", note d'analyse n°112, France stratégie, septembre 2022. lien : https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/fs-2022-na-112-inegalites_environnementales-septembre_0.pdf
17- Marcel Goldberg, Mireille Coeuret-Pellicer, Céline Ribet et Marie Zins, "Cohortes épidémiologiques et bases de données d'origine administrative : Un rapprochement potentiellement fructueux", Med Sci (Paris), 2012 ; 28 pp. 430–434
18- Lukacs B, Cornu JN, Aout M, Tessier N, Hodée C, Haab F, Cussenot O, Merlière Y, Moysan V, Vicaut E., "Management of lower urinary tract symptoms related to benign prostatic hyperplasia in real-life practice in France: a comprehensive population study."Eur Urol, septembre 2013(3) pp.493-501. lien : https://pubmed.ncbi.nlm.nih.gov/23465519/
19- Morgane Le Bail, Zeynep Or (dir.), "Atlas des variations de pratiques médicales. Recours à dix interventions chirurgicales, Edition 2016", Irdes, novembre 2016. lien : https://www.irdes.fr/recherche/ouvrages/002-atlas-des-variations-de-pratiques-medicales-recours-a-dix-interventions-chirurgicales.pdf
20- Jan Piasecki, Phaik Yeong Cheah, "Ownership of individual-level health data, data sharing, and data governance"_ *BMC Med Ethics*, octobre 2022. lien : doi: 10.1186/s12910-022-00848-y. PMID: 36309719
21- "Livre blanc préliminaire du conseil scientifique du CNRS sur les entraves administratives à la recherche", Conseil scientifique du CNRS, mai 2023. lien : https://www.cnrs.fr/comitenational/cs/recommandations/Rapport_Entraves_vf.pdf
22- "Espace européen des données de santé", Commission européenne, 2023. lien : https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_fr (consulté le 2 octobre 2023).
23- "Advancing data sharing to improve health for all in europe, Main findings of joint action Towards the European Health Data Space 2021–2023", Sitra, Tehdas, Markus Kalliola, Elina Drakvik and Maria Nurmi (Ed.), 2023. lien : https://tehdas.eu/results/eu-wide-collaboration-needed-to-optimise-health-data-use-for-research-and-innovation/
24- Baptiste Couvy-Duchesne, Simona Bottani, Etienne Camenen, Fang Fang, Mulusew Fikere, Juliana Gonzalez-Astudillo, Joshua Harvey, Ravi Hassanaly, Irfahan Kassam, Penelope A. Lind, Qianwei Liu, Yi Lu, Marta Nabais, Thibault Rolland, Julia Sidorenko, Lachlan Strike, Margie Wright, "Main existing datasets for open data research on humans", Machine Learning for Brain Disorders book, 2023, pp 753-806. lien : https://link.springer.com/book/10.1007/978-1-0716-3195-9
25- " Il faut faire des données de santé un bien commun pour la recherche", *Le Monde*, 31 mars 2023. lien : https://www.lemonde.fr/idees/article/2023/03/31/intelligence-artificielle-il-faut-faire-des-donnees-de-sante-un-bien-commun-pour-la-recherche_6167803_3232.html
26- Kenneth P. Seastedt, Patrick Schwab, Zach O'Brien, Edith Wakida, Karen Herrera, Portia Grace F. Marcelo, Louis Agha-Mir-Salim, Xavier Borrat Frigola, Emily Boardman Ndulue, Alvin Marcelo, and Leo Anthony Celi, "Global healthcare fairness: We should be sharing more, not less, data", PLOS Digit Health, octobre 2022
27- "Données de santé : une réforme encore en cours de chargement", Rapport d'information n° 873, Commission des affaires sociales, Sénat, 12 juillet 2023. lien : https://www.senat.fr/rap/r22-873/r22-8731.pdf
28- Yann Joly, Stephanie O.M. Dyke, Bartha M. Knoppers, Tomi Pastinen, "Are Data Sharing and Privacy Protection Mutually Exclusive?", Cell, n°167, 17 novembre 2016

HEALTH DATA HUB

Table des matières