

Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé



Rapport

Jérôme MARCHAND-ARVIER

Pr Stéphanie ALLASSONNIERE

Aymeril HOANG

Dr Anne-Sophie JANNOT

Rapporteurs

Emilie FAUCHIER-MAGNAN (IGAS), Julien FRADEL (Conseil d'État)

5 décembre 2023

PRÉAMBULE

Ce préambule est une introduction personnelle, que je souhaite d'autant plus écrire que mes nouvelles fonctions ne me permettront pas de présenter publiquement les recommandations de ce rapport, ce que les autres membres de la mission feront parfaitement.

La mission qui nous a été confiée consistait à jeter les bases pour la construction d'une nouvelle feuille de route pour l'utilisation secondaire des données de santé. A ce titre, la lettre de mission n'abordait pas le sujet de l'hébergement des infrastructures nécessaires pour l'utilisation secondaire des données de santé, et en particulier le choix de la plateforme technologique d'hébergement du Health Data Hub. Très vite, nous avons constaté que ce sujet, s'il n'était pas nécessairement le plus important, revenait en permanence dans les échanges et constituait un point de blocage, réel ou supposé, pour, d'une part, faciliter et accélérer l'accès aux données de la base principale du système national des données de santé (SNDS), d'autre part, engager l'ensemble des acteurs dans une logique coopérative de partage des données de santé. C'est pourquoi la mission a rapidement décidé qu'il était nécessaire de traiter ce sujet et même d'en faire un préalable. Je me permets d'insister sur le terme « préalable », qui implique que traiter ce sujet est nécessaire mais pas suffisant. L'ensemble des autres recommandations de la mission nous paraissent bien plus importantes que ce seul sujet. Mais elles n'atteindront leur plein potentiel que si ce sujet a pu être traité clairement et rapidement.

Le rapport assume une orientation résolument optimiste. Il ne s'agit pas d'un optimisme naïf : beaucoup de progrès ont été accomplis en quelques années en matière d'utilisation secondaire des données de santé. L'ensemble des acteurs ont conscience de l'importance des données de santé. Tous savent aussi qu'il est nécessaire de bien structurer, avec des infrastructures adaptées, ces données de santé pour pouvoir les réutiliser d'une manière optimale, et des investissements notables ont été réalisés en la matière. Tous savent enfin, et l'assument désormais, qu'une logique d'ouverture et de partage de ces données fait progresser la science, la recherche et la santé. Mais cette orientation relève également d'un optimisme de la volonté, car force est de reconnaître que ces progrès ne peuvent masquer le « verre à moitié vide », avec certes des processus d'accès aux données longs, complexes et coûteux, mais aussi une culture collective qui, trop souvent encore, conduit *in fine* à ne pas tout faire ensemble pour que concrètement, le partage des données de santé soit plus fluide et plus rapide. Il en va certes de la responsabilité du système – et donc de la nécessité de procéder à des évolutions de ce système –, mais il en va aussi de la responsabilité de chacun des acteurs, qui doivent, avec confiance, s'engager résolument dans une logique et une culture d'ouverture et de partage qui n'est pas encore suffisamment présente.

Dans ce cadre, la mission a privilégié autant que possible des recommandations dont la mise en œuvre est possible à courte échéance. Pour celles qui nécessitent des délais plus longs – notamment en cas d'évolution normative nécessaire –, il est souvent envisageable de préparer ces évolutions voire de les anticiper dans les faits. Dans cette logique d'avoir des recommandations les plus opérationnelles possibles, une partie de ces orientations ont été préparées dans le cadre de trois ateliers, en novembre 2023, avec les principaux acteurs concernés. Je suis, comme tous les membres de la mission, ressorti de ces ateliers avec une grande confiance sur la volonté et la

capacité des acteurs à coopérer ensemble, sous l'impulsion politique nécessaire, pour mettre en œuvre rapidement et efficacement les recommandations qui seraient endossées.

En synthèse le potentiel est immense. La bonne volonté est présente. Mais la confiance est absolument nécessaire, tout comme est nécessaire un engagement constant des pouvoirs publics, afin que ce sujet de l'utilisation secondaire des données de santé soit piloté comme il se doit, avec détermination et régularité.

Je voudrais conclure par quelques remerciements.

Remerciements tout d'abord qui vont à l'ensemble des personnes avec qui nous avons échangé dans le cadre de la mission. Compte tenu des délais impartis, et malgré de très nombreux échanges, il ne nous a pas été possible d'échanger avec toutes celles et ceux qui auraient pu apporter des éléments utiles à la mission. Malgré ce regret, chacun des échanges que nous avons eus nous a permis de mesurer l'excellence des acteurs et leur plein engagement pour l'amélioration des soins, pour la recherche et la création de valeur.

Remerciements ensuite aux membres de la mission, Stéphanie Allassonnière, Anne-Sophie Jannot et Aymeril Hoang. Nous avons pendant plusieurs mois travaillé main dans la main, et l'engagement des membres de la mission, malgré leurs nombreuses autres fonctions, a été total. Les différences des parcours, des expertises et des convictions des membres de la mission ont constitué un formidable enrichissement et, au final, une complémentarité très précieuse pour mener à bien cette mission.

Remerciements enfin aux deux rapporteurs de la mission, Emilie Fauchier-Magnan, inspectrice des affaires sociales – et je remercie à ce titre l'IGAS d'avoir apporté son concours – et Julien Fradel, maître des requêtes en service extraordinaire au Conseil d'Etat. Sans leur appui décisif et constant, la mission n'aurait pas pu être menée à bien, et ce rapport leur doit naturellement beaucoup.

Jérôme Marchand-Arvier, le 5 décembre 2023

SYNTHÈSE

Par lettre en date du 31 mai 2023, le ministre de l'Économie, des finances et de la souveraineté industrielle et numérique, la ministre de l'enseignement supérieur et de la recherche et le ministre de la santé et de la prévention ont lancé une **mission visant à poser les bases d'une feuille de route en matière de réutilisation des données de santé**. Son animation a été confiée à Jérôme Marchand-Arvier. La mission était également composée de Stéphanie Allassonnière, Anne-Sophie Jannot et Aymeril Hoang¹.

L'utilisation primaire des données de santé renvoie à la collecte de données à l'occasion de la prise en charge des patients, quand l'usage secondaire désigne l'utilisation de ces données pour d'autres finalités comme la recherche et l'innovation ou le pilotage ou encore l'amélioration de la qualité des soins. Ainsi, la conception d'algorithmes pour mieux détecter des pathologies, la réalisation d'études pour mesurer les effets en vie réelle de médicaments ou encore l'élaboration de tableaux de bord pour piloter une crise sanitaire s'appuient sur des usages secondaires des données de santé. La mission a par ailleurs retenu la définition large des données de santé adoptée par le règlement européen sur la protection des données personnelles (RGPD), intégrant dans son périmètre les données médicales, les données médico-administratives mais aussi les données de recherche en santé.

La mission a tout d'abord dressé un état des lieux de l'utilisation secondaire des bases de données en santé. Elle émet ensuite des recommandations destinées à alimenter une stratégie ambitieuse d'utilisation secondaire des données de santé, permettant d'accélérer notablement l'accès aux données de santé en France, de favoriser le déploiement de l'intelligence artificielle et de contribuer à l'amélioration de la compétitivité française en termes de recherche et d'innovation.

Le potentiel important de réutilisation du riche patrimoine de données de santé français est encore sous-exploité, dans un contexte international compétitif.

A l'échelle mondiale, la réutilisation des données de santé connaît un fort dynamisme, avec le développement d'un tissu important d'entreprises spécialisées dans la mobilisation des données de santé et dans la génération de valeur à partir de ces données. La crise de la covid-19 a conduit à une accélération importante sur ce champ, du fait de l'augmentation du nombre d'essais cliniques, de la demande accrue de développement de nouveaux médicaments et vaccins et des besoins continus d'évaluation et réévaluation de ces produits. L'intelligence artificielle (et notamment l'intelligence artificielle générative) offre également des perspectives nouvelles prometteuses, par exemple dans la mise au point de nouvelles modalités d'essais cliniques avec des bras partiellement ou complètement virtuels ou de nouvelles approches thérapeutiques avec la création de jumeaux numériques.

Le potentiel d'exploitation des données de santé reste très important aussi pour les politiques publiques, que ce soit pour la prévention en santé, l'amélioration de la qualité des soins, le

¹ La mission a bénéficié du concours d'Emilie Fauchier-Magnan, inspectrice des affaires sociales, et de Julien Fradel, maître des requêtes en service extraordinaire au Conseil d'Etat.

financement des produits de santé ou encore la surveillance sanitaire et la politique de santé publique. A titre d'exemple, l'utilisation de données de vie réelle constitue un potentiel important pour l'évaluation des produits de santé, même si cela nécessite une évolution des doctrines d'évaluation des autorités.

Dans ce contexte, le positionnement de la France apparaît relativement en retrait par rapport à d'autres pays étrangers, un certain nombre d'entreprises et de chercheurs français privilégiant l'utilisation de données à l'étranger, auprès de pays très compétitifs dans ce domaine (notamment les États-Unis, Israël et le Royaume-Uni). Le nombre de publications liées aux bases de données françaises demeure ainsi très largement inférieur à certaines bases étrangères très accessibles.

La France dispose pourtant **d'un patrimoine de données de santé très riche, dont l'utilisation est croissante, mais qui demeure encore cloisonné et peu interopérable** :

- la base principale du système national des données de santé (SNDS), qui regroupe notamment les données médico-administratives de l'assurance maladie et les données de facturation des établissements de santé, constitue une base internationalement reconnue du fait de son exhaustivité, de sa complétude et de la fiabilité des données contenues. Son potentiel d'exploitation, qui a beaucoup progressé depuis l'ouverture de l'accès à ses données en 2016 et la création de nombreux accès permanents très appréciés des chercheurs, peut encore être décuplé par des appariements plus fréquents et plus rapides avec des données médicales ou avec des données d'autres champs sectoriels (données socio-fiscales, environnementales, etc.) ;
- le développement de nombreux entrepôts de données de santé (EDS), principalement dans les établissements de santé, constitue un progrès récent et important. Ils regroupent des données médicales nombreuses et produites en routine. Un financement public, notamment dans le cadre du plan d'investissement France 2030, a contribué à leur essor récent, mais cet accompagnement financier n'est toutefois pas pérenne. La qualité des données reste toutefois souvent difficile à connaître et très hétérogène, avec un manque d'harmonisation de leur structuration et l'absence d'utilisation de standards de données partagés ;
- les cohortes et registres sont nombreux en France, mais particulièrement hétérogènes et insuffisamment structurés. Ils contiennent des données médicales souvent riches et très utiles pour la recherche scientifique. Ces bases de données restent toutefois pour beaucoup isolées, et mériteraient d'être davantage croisées avec d'autres bases de données pour enrichir leur contenu. Leur financement, rarement pérenne, reste en retrait par rapport à certains pays étrangers.

De nombreuses autres bases de données (autres bases du SNDS, bases épidémiologiques, études, etc.) complètent un paysage des bases de données en santé riche, mais peu lisible. Par ailleurs, les bases de données françaises restent peu appariées entre elles, alors que les analyses des parcours de vie et de soin complets apportent une plus-value importante, notamment dans une optique de prévention en santé.

Les freins à la réutilisation des données de santé restent nombreux.

Plusieurs facteurs concourent à ce que l'accès aux données de santé en France soit globalement long et complexe, pour les données de la base principale du SNDS comme pour les autres bases de données.

L'éparpillement des bases de données, leur contenu hétérogène et leur documentation insuffisante constituent un premier frein à la réutilisation des données par les acteurs économiques comme par les chercheurs. En dépit de premières initiatives, il manque encore une cartographie transverse des bases de données, à jour, susceptible de renvoyer vers des descriptions homogènes du contenu précis des bases de données et de leurs métadonnées.

La complexité et la longueur des démarches réglementaires constituent un second frein pour l'accès aux données. La France est un des rares pays, au niveau européen, à avoir maintenu un régime d'autorisation préalable par une agence de protection des données pour la recherche et les études nécessitant un accès aux données de santé, obligatoire en cas d'appariement entre plusieurs bases. Sont ainsi nécessaires pour mener un projet :

- un avis d'un comité scientifique et éthique (pour les recherches n'impliquant pas la personne humaine, le Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé - CESREES)
- une autorisation de la Commission nationale de l'informatique et des libertés (CNIL).

En dépit de l'accompagnement réalisé par la Plateforme des données de santé (par la suite du rapport désignée Health data hub), qui joue depuis 2019 un rôle de guichet unique pour ces procédures d'avis, et d'une maîtrise globale des délais par le CESREES et la CNIL, la procédure reste perçue comme longue et complexe, d'autant qu'elle peut se rajouter à l'examen par un comité scientifique et éthique d'un EDS ou d'une cohorte.

Il existe certes des procédures simplifiées permettant de se dispenser de l'autorisation de la CNIL, voire de l'avis du CESREES : accès permanents aux données de la base principale du SNDS, méthodologies de référence, référentiels, décisions uniques, etc. Si leur existence est appréciée, elles ne suffisent toutefois pas à fluidifier le système, du fait de la complexité du cadre général réglementaire, d'une appropriation insuffisante par une partie des acteurs et d'une méthodologie de construction qui associe insuffisamment les acteurs.

De même, les exigences en matière d'information des personnes contribuent à la complexité de la procédure réglementaire, le RGPD posant le principe d'une ré-information individuelle des patients à chaque réutilisation de leurs données. L'effort de transparence et de pédagogie vis-à-vis des citoyens sur la réutilisation de leurs données demeure par ailleurs insuffisant.

Enfin, la longueur de l'étape de contractualisation pour l'accès aux données de santé constitue une difficulté majeure pour l'accès aux données de santé.

S'agissant de l'accès aux données de la base principale du SNDS, **l'étape de contractualisation et de mise à disposition des données par la caisse nationale de l'assurance maladie (CNAM) est particulièrement longue**, avec un délai moyen d'accès aux données de 10 à 12 mois, en hausse (+ 56 % du délai médian entre 2020 et 2022), hausse certes en partie expliquée par une

augmentation importante des demandes (35 projets livrés par la CNAM en 2018 contre 165 en 2023) et par leur complexité croissante. Ce délai élevé est très problématique pour de nombreux acteurs publics comme privés, conduisant à des renoncements de la part d'un certain nombre d'acteurs. S'il est prévu depuis 2019 que le Health data hub prenne le relai de la CNAM pour la mise à disposition de ces données, cette transition n'a toujours pas été opérée en raison du blocage du transfert des données de la base principale du SNDS, lié au choix de la solution logicielle Azure de Microsoft pour la plateforme technologique du Health data hub².

Pour les autres bases de données que la base principale du SNDS, l'étape de contractualisation peut aussi être très longue, du fait notamment de modèles de tarification hétérogènes, peu lisibles et parfois peu compétitifs. L'absence de modalités de financement pérenne des bases de données contribue aux exigences élevées des producteurs de données.

Malgré l'investissement de la puissance publique pour la réutilisation des données de santé et les réalisations volontaristes du Health data hub, l'écosystème est marqué par un déficit de coopération et de confiance.

La création du Health data hub en 2019 a témoigné d'un investissement public important pour faire progresser la réutilisation des données de santé et structurer l'écosystème. Dans le même temps, des financements importants ont d'ailleurs été consacrés aux projets numériques en santé, en particulier dans le cadre du plan d'investissement France 2030.

Même si le Health data hub ne peut remplir une de ses missions centrales, la mise à disposition des données de la base principale du SNDS, son action a permis de premières réalisations encourageantes, à l'échelle française comme à l'échelle internationale. Il s'est imposé comme un partenaire central de l'écosystème des données de santé, avec un accompagnement des porteurs de projets apprécié et un rôle de sensibilisation sur l'utilisation des données de santé. Ses capacités techniques sont également reconnues.

Cependant, **la mise en œuvre du catalogue du SNDS par le Health data hub fait l'objet de crispations qui ne se limitent pas au sujet de l'hébergement de sa plateforme technique**. Ce catalogue - notion qui suscite une certaine confusion dans l'esprit des acteurs - désigne les bases de données du SNDS, en dehors de la base principale, qui ne couvrent pas l'ensemble de la population. Sa mise en œuvre opérationnelle suppose actuellement la réalisation d'une copie des données de ces bases sur la plateforme du Health data hub³. Cette perspective de centralisation systématique a été mal accueillie par une large part des producteurs de bases de données, et se traduit par un mouvement de défiance d'un certain nombre d'acteurs vis-à-vis du Health data hub.

Par ailleurs, **le foisonnement des initiatives à l'œuvre en matière de réutilisation des données de santé se traduit par un manque d'articulation et de coordination interministérielles**. A titre d'exemple, plusieurs initiatives coexistent en matière de référencement des bases de données en santé (portail Fresh porté par l'Inserm, répertoire du Health data hub). De même, l'articulation

² La demande d'autorisation auprès de la CNIL relative à l'hébergement du SNDS a finalement été retirée par la plateforme en janvier 2022 compte tenu des risques juridiques liés à une solution d'hébergement états-unienne et compte tenu de la forte sensibilité politique de la question.

³ Article R 1461-3 du code de la santé publique

entre la plateforme du Health data hub, l'outil France Cohortes et la plateforme de l'Institut national du cancer mériterait d'être clarifiée.

Or le comité stratégique des données de santé, créé par la loi relative à l'organisation et à la transformation du système de santé du 24 juillet 2019, ne joue pas suffisamment ce rôle d'impulsion et de coordination de la politique d'utilisation secondaire des données de santé, même s'il a réalisé de premiers travaux de coordination intéressants, s'agissant notamment des entrepôts de données de santé (socle de données, grilles de tarification, etc.).

De manière générale, même si l'état d'esprit a progressé, **persistent un déficit de confiance de la part des producteurs de données et des réticences au partage des données**, liées en particulier à une méconnaissance des règles de partage, à des craintes d'absence de reconnaissance scientifique ou économique, et plus généralement au maintien de « réflexes propriétaires », avec la volonté de maintenir une forme de suivi ou de contrôle sur l'utilisation de « leurs » données.

Ainsi, si tous les acteurs publics et privés louent le principe d'un partage des données de santé à des fins de recherche et d'innovation, et si chaque acteur pris isolément poursuit des objectifs légitimes (pertinence de la méthodologie de réutilisation des données, efficacité ou sécurité de l'architecture informatique, respect de la réglementation, ...), cela se traduit par un environnement globalement peu coopératif et qui peut décourager les porteurs de projets.

La mission recommande en préalable d'arbitrer la question de l'hébergement de la copie de la base principale du SNDS par le Health data hub

La mission a décidé de traiter comme un préalable le sujet de l'hébergement de la plateforme du HDH et d'une copie de la base principale du SNDS, partant du constat que ce sujet, au-delà de son impact réel sur la capacité du HDH à assurer pleinement sa mission et sur le formidable potentiel d'exploitation des données de la base principale du SNDS, entraînait un blocage plus large et risquait de devenir le symbole d'un échec.

La mission n'entend pas revenir sur le choix de cette solution et les raisons qui l'ont sous-tendu, ni sur le débat juridique concernant sa conformité avec l'exigence de protection des données personnelles contre les lois extraterritoriales américaines. Elle constate que la migration de la plateforme du HDH sur un cloud souverain a déjà été actée par le gouvernement. **Il s'agit désormais d'acter publiquement et officiellement cette orientation, d'en préciser les modalités et d'en tirer toutes les conséquences**, en particulier s'agissant de la période transitoire, le temps qu'une solution d'hébergement pleinement satisfaisante, en termes de sécurité et de souveraineté, soit effectivement disponible.

Les informations réunies par la mission permettent d'estimer **qu'un horizon de 24 mois est ambitieux mais crédible pour l'hébergement du HDH sur un cloud qualifié « SecNumCloud »**. La réussite d'une telle opération, dans un tel calendrier, nécessite une forte mobilisation des acteurs du cloud, mais également un pilotage vigoureux de cette opération par le HDH et les acteurs publics impliqués, en particulier de la Dinum au titre de ses responsabilités dans l'animation de la politique industrielle de l'État en matière de cloud souverain.

Un arbitrage immédiat sur la solution transitoire est aussi nécessaire. Plusieurs options sont détaillées dans le rapport afin de permettre au HDH d'enfin piloter la gestion et la mise à disposition d'une copie de la base principale du SNDS.

La mission formule des recommandations pour alléger la réglementation et accélérer la mise à disposition des données de la base principale du SNDS, et plus généralement des données de santé.

Une première série de recommandations concerne les procédures simplifiées d'accès aux données de santé. Le mode d'élaboration de ces procédures devrait être réformé pour renforcer la co-construction avec les acteurs de la réutilisation des données, en confiant un rôle plus important au Health data hub et au Comité stratégique des données de santé dans son élaboration. Le champ de ces procédures simplifiées doit être étendu, et leur architecture allégée, afin que l'autorisation devienne l'exception.

La mission recommande aussi de rénover la procédure d'examen éthique et scientifique pour les études et recherches n'impliquant pas la personne humaine, dans une logique de subsidiarité vis-à-vis des comités éthiques et scientifiques locaux, dès lors qu'ils respecteront un cahier des charges national.

Il est aussi nécessaire d'anticiper le futur règlement sur l'espace européen des données de santé (dit règlement EHDS), alors que ses négociations sont encore en cours. La préservation des mécanismes d'accès permanent et de procédures simplifiées constitue un premier enjeu, tout comme la capacité à tarifier les redevances d'accès aux données pour couvrir l'ensemble des coûts supportés par les producteurs de données. La rénovation des mécanismes d'information individuelle doit aussi permettre d'alléger la réglementation, en contrepartie d'un effort de transparence accrue auprès du public et d'une facilitation de l'exercice des droits individuels concernant les données de santé, en particulier le droit d'opposition.

D'autres mesures visent à accélérer la mise à disposition des données de la base principale du SNDS. Au-delà de la question de l'hébergement des données déjà évoquée, **la piste la plus prometteuse est celle d'une responsabilisation accrue des responsables de traitement en termes de minimisation des données,** dans l'esprit du RGPD. Ainsi, les porteurs de projet auraient accès, de manière temporaire, à une copie intégrale de la base principale, pour réaliser l'extraction des données selon les principes le cas échéant validés par le CESREES et la CNIL. L'étape d'extraction serait ainsi considérablement accélérée. Cette option, qui emporte le consensus des principaux acteurs concernés, doit être mise en place sans tarder pour les acteurs ayant une bonne connaissance de la base principale du SNDS.

La publication d'indicateurs sur les délais de mise à disposition des données de la base principale du SNDS sera quant à elle de nature à favoriser la transparence du système. Elle gagnerait à être établie de manière concertée pour être élargie à l'ensemble des bases de données en santé.

Enfin, le développement de formations plus nombreuses et décentralisées à l'utilisation des données de la base principale du SNDS, ainsi que le développement d'outils de « bac à sable »

pour confirmer la faisabilité de projets d'études ou de recherche constituent d'autres leviers pour faciliter l'accès aux données pour les chercheurs.

Les missions et le positionnement du Health data hub devront être refondés, et le pilotage stratégique interministériel renforcé.

Le positionnement du Health data hub comme véritable pilote opérationnel de l'écosystème des données de santé suppose de clarifier au préalable les obligations liées à l'inscription de bases de données au catalogue du SNDS. L'intérêt d'un catalogue de collections de données à caractère national ou local pour « faire monter » en visibilité les bases concernées et favoriser leur mobilisation à l'échelle nationale ou internationale est indéniable. En revanche, la réplcation de chacune des bases sur la plateforme technologique du HDH ne saurait continuer à conditionner leur inscription au catalogue du SNDS, car c'est un frein à l'enrichissement de ce catalogue. Dès lors, la mission recommande de faire du stockage d'une copie des données des bases inscrites au catalogue du SNDS une faculté et non plus une obligation.

Le Health data hub pourra et devra ensuite jouer un rôle renforcé et fédérateur, avec un rôle de garant du partage des données. Celui-ci se traduirait par un rôle d'impulsion et d'animation de principes de partage de données, de mise en visibilité des données de santé, mais aussi de définition de standards et de normes d'interopérabilité, en associant l'Agence du numérique en santé. Le HDH pourrait aussi assumer de manière plus importante la mission d'accompagner les porteurs de projets dans leurs démarches réglementaires. Si le HDH conserverait une plateforme technique permettant la mise à disposition d'espaces de travail pour des projets exploitant des données, la mission s'interroge en revanche sur le maintien de son rôle dans le financement de projets innovants, le HDH devant jouer un rôle plus proche d'un régulateur sur plusieurs aspects, notamment dans la perspective du futur règlement européen EHDS. La mission recommande en effet de désigner le HDH comme unique organisme responsable de l'accès aux données de santé (HDAB), tout en prévoyant le maintien d'un avis conforme de la CNIL pour l'octroi de certaines autorisations d'accès sensibles.

L'exercice de ces nouvelles missions par le Health data hub suppose **que le comité stratégique des données de santé joue un rôle de pilotage stratégique renforcé** en assurant l'articulation et la cohérence entre les multiples initiatives en matière de structuration de l'écosystème des bases de données. A ce titre, **la mission recommande de faire évoluer la gouvernance et les missions du comité stratégique des données de santé pour en faire une réelle instance d'impulsion de la politique de réutilisation des données de santé.** La co-présidence du comité par le ministre chargé de la santé et le ministre chargé de la recherche s'impose pour garantir la dimension réellement interministérielle de cette instance. La création en son sein d'un comité de pilotage plus restreint facilitera également le caractère stratégique du pilotage.

D'autres leviers devront être mobilisés pour renforcer l'ambition de la politique de la réutilisation des données de santé.

La réalisation d'un référencement des bases de données en santé est indispensable pour renforcer la lisibilité du paysage et maximiser l'ouverture et la mobilisation de ces bases. Un cadre national doit permettre de mieux articuler les initiatives existantes en matière de référencement, de définir

le contenu des différents niveaux de cartographie et d'harmoniser les différentes métadonnées produites.

Des standards de données doivent aussi être adoptés pour renforcer l'interopérabilité des bases de données, en particulier s'agissant des entrepôts de données de santé. Pilotés de manière opérationnelle par le Health data hub, ces travaux devraient se traduire par des normes d'interopérabilité contraignantes.

Par ailleurs, **le partage des données sera d'autant plus facilité que le modèle économique et les règles de reconnaissance seront clarifiés**. Des financements pérennes et suffisants des infrastructures de données de santé (notamment les entrepôts de données de santé et les cohortes) sont ainsi nécessaires. Les travaux pour établir des grilles de tarification harmonisées entre les bases de données doivent aussi être accélérés, et des contrats-types devraient être établis pour réduire les délais de négociation des contrats. Enfin, la définition concertée de principes de reconnaissance scientifique, notamment en termes de durée d'éventuel embargo et de règles de publication, renforcera la confiance des producteurs de données pour le partage.

Plusieurs préconisations visent à renforcer la transparence vis-à-vis des patients sur la réutilisation des données de santé les concernant. La mission recommande aussi d'étudier une simplification des conditions d'exercice des droits « d'opt out » des patients qui s'opposent à la réutilisation de leurs données.

Enfin, la mission formule différentes recommandations pour poursuivre l'enrichissement du patrimoine de données en santé, à travers de nouvelles bases de données sur certains champs manquants ou insuffisamment explorés, la consolidation de la base principale du SNDS, ou des croisements intersectoriels susceptibles d'enrichir la recherche en santé, en particulier avec les champs environnementaux et sociaux. Une réflexion sur une réutilisation, à terme, des données de Mon espace santé pour la recherche publique mériterait aussi d'être initiée.

RECOMMANDATIONS DE LA MISSION

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
Préalable : arbitrer la question de l'hébergement de la copie de la base principale du SNDS par le HDH			
1.	Programmer l'arrêt de l'hébergement sur Azure du HDH et lancer les travaux pour l'hébergement du HDH, y compris la copie de la base principale du SNDS, sur un cloud SecNumCloud, à horizon de 24 mois, échéance ambitieuse mais crédible à ce stade.	HDH, en lien avec le ministère de la santé et la Dinum	
2.	Définir, dans les six mois, et sous pilotage du HDH, une solution transitoire pour accélérer notablement la mise à disposition des données, en particulier par la mise à disposition d'une copie de la base principale du SNDS.	HDH/ DNS	
3.	Procéder à un audit « flash » pour définir la solution transitoire souveraine la plus adaptée en l'état actuel, en incluant l'examen du centre d'accès sécurisé aux données (CASD) dans cet audit, afin de permettre le transfert d'une copie de la base principale du SNDS (cf. recommandation n°2).	HDH/DNS, en lien avec la Dinum	
Alléger les processus réglementaires			
4.	Mobiliser les leviers de négociation disponibles pour ajuster le projet de règlement EHDS, notamment lors de son examen par le Parlement européen, afin de préserver les mécanismes d'accès permanent et de procédures simplifiées ; anticiper l'éventualité d'une mise en cause potentielle de ces mécanismes en élaborant de nouvelles modalités d'accès simplifié, avec des effets équivalents.	DNS/ COSTRAT/ CNIL	Règlement européen EHDS
5.	Changer le mode d'élaboration des méthodologies de référence, en prévoyant leur adoption par arrêté du ministre chargé de la santé après avis de la CNIL, en confiant au HDH la mission de recueillir les attentes de l'écosystème en matière d'évolution de ces méthodologies et au Comité stratégique des données de santé la mission de proposer des projets de référentiels simplifiés et de méthodologies de référence.	Ministère de la santé	Législatif (art 66 LIL ; art L. 1462-1 CSP)
6.	Rationaliser le paysage des méthodologies de référence en privilégiant le format des méthodologies de référence "cadres".	COSTRAT/ HDH/CNIL	

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
7.	Étendre le champ des méthodologies de référence pour notamment permettre le recueil et le traitement du NIR et le chaînage de données issues de plusieurs bases dans le cadre de projets de recherche.	COSTRAT/ CNIL/HDH	
8.	Fixer un objectif quantifié de diminution du nombre de projets nécessitant une autorisation de la CNIL pour que, de manière effective, l'autorisation devienne l'exception par rapport aux procédures simplifiées.	CNIL/ COSTRAT/ HDH	
9.	Clarifier, si besoin dans le projet de règlement EHDS, que, par dérogation à l'obligation du RGPD de ré-information individuelle du patient à chaque réutilisation de ses données de santé, le patient, qui a déjà été informé, à l'occasion de la collecte de ses données, de la possibilité que ses données soient réutilisées, n'a pas à être à nouveau informé.	DNS	Règlement européen EHDS
10.	Renforcer la formation des porteurs de projet, des équipes juridiques et des délégués à la protection des données (DPO) à la réglementation applicable en matière de recherche en santé et mobiliser le HDH pour les conseiller sur la procédure réglementaire la plus adaptée au projet.	HDH/CNIL	
11.	Dispenser les projets de recherche n'impliquant pas la personne humaine d'un avis du CESREES, dès lors qu'un comité scientifique et éthique local respectant un cahier des charges national a rendu un avis.	SGMAS	Législatif (art 76 LIL)
Accélérer la mise à disposition des données du SNDS			
12.	Avancer collectivement et rapidement sur la délégation au porteur de projet de la responsabilité de la minimisation, avec une mise à disposition temporaire d'une copie du SNDS, afin d'accélérer significativement l'accès aux données du SNDS.	CNAM/ HDH/CNIL	
13.	Développer à brève échéance un indicateur de délai sur la mise à disposition des données du SNDS ; à moyen terme, étendre cette logique de transparence à l'ensemble des bases de données en santé avec l'élaboration concertée d'indicateurs de délai au sein du comité stratégique des données de santé.	CNAM/ COSTRAT/ HDH	

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
14.	Développer les formations à l'utilisation de la base principale du SNDS, dans une logique de démultiplication et de « décentralisation », avec un financement par la stratégie d'accélération du numérique en santé.	HDH/ Org de recherche/ Universités/ DNS	
15.	Développer des outils de « bac à sable » accessibles librement ou de manière simplifiée, permettant de confirmer la faisabilité de projets d'études ou de recherche sur les données de la base principale du SNDS et, si possible, sur d'autres bases de données de santé.	CNAM/ HDH	
16.	Expérimenter la parallélisation de l'examen des demandes d'autorisation de traitement de données de santé par le CESREES (ou, le cas échéant, par le CSE de la base de données requêtée) et par la CNIL, en commençant par les demandes d'appariement avec la base principale du SNDS.	CESREES/ CNIL	Dérogation expérimentale au décret n° 2019-536 (art. 91- 94)
Refonder le positionnement du Health data hub			
17.	Faciliter l'enrichissement du catalogue du SNDS en allégeant le processus d'inscription et en faisant du versement d'une copie d'une base de données sur la plateforme du HDH une faculté et non une obligation, sous réserve du respect des principes de partage et des standards d'interopérabilité par les bases inscrites au catalogue.	Ministère de la santé/ HDH	Réglementaire (Art R. 1461-3 CSP)
18.	Redéfinir et renforcer les missions du HDH afin d'en faire un acteur fédérateur de l'écosystème de l'utilisation secondaire des données de santé.	Ministère de la santé / HDH	Législatif (Art L. 1462-1 CSP)
19.	Charger le comité stratégique des données de santé de veiller à l'articulation et à la mise en cohérence des différentes initiatives de structuration de l'écosystème des bases de données de santé, en s'appuyant sur le HDH pour coordonner les initiatives.	COSTRAT	
20.	Désigner le HDH comme HDAB au sens du règlement EHDS, tout en prévoyant que l'octroi par le HDAB des autorisations pour les accès ne relevant pas d'un référentiel ou d'une méthodologie de référence est accordé par le HDH après avis conforme de la CNIL.	Ministère de la santé/ HDH/ CNIL	Législatif (art 76 LIL, art L. 1462-1 CSP)

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
Renforcer le pilotage stratégique et consolider le modèle économique			
21.	Rénover la gouvernance et les missions du comité stratégique des données de santé pour en faire une structure interministérielle d'impulsion de la politique d'utilisation secondaire des données ; renforcer les ressources dédiées à l'animation du comité.	Ministère chargé de la santé/ Ministère de la recherche	Réglementaire (R. 1461-10 du CSP et arrêté 29/06/21)
22.	Définir un cadre national pour mettre en visibilité les bases de données de santé, en clarifiant l'articulation entre les différents projets en cours ; en confier le pilotage au comité stratégique des données de santé.	COSTRAT	
23.	Définir sous l'égide du Health data hub, en associant étroitement l'agence du numérique en santé et en concertation avec les acteurs, des normes d'interopérabilité pour les entrepôts de données hospitaliers ; intégrer dans ces normes une exigence de partage des schémas de structuration des données par les éditeurs de logiciels.	HDH/ANS En lien avec la DNS	Réglementaire
24.	Établir des financements pérennes et suffisants des infrastructures de données de santé (entrepôts de données de santé hospitaliers, registres, cohortes) et du Health data hub pour favoriser la logique de partage des données.	DSS/ DGOS/ DGS	Lois de finances (LFSS/LF)
25.	Préparer la mise en place d'une redevance pour l'accès aux données de la base principale du SNDS dès lors que les délais d'accès le permettront	COSTRAT/ CNAM/ HDH/	
26.	Approfondir et accélérer les travaux pour établir des grilles tarifaires harmonisées pour le financement par redevances des bases de données en santé et instaurer une obligation de transparence sur les modalités de tarification des producteurs de données.	COSTRAT/ HDH	
27.	Définir des modèles de contrats-type pour l'accès aux données de santé, en commençant par le modèle de tarification par redevances.	COSTRAT/ HDH	
28.	Etablir des principes de reconnaissance scientifique pour la réutilisation de données (durée d'embargo, règles de publication...) pour favoriser le partage	COSTRAT/ HDH	

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
Renforcer la transparence vis-à-vis des patients sur l'utilisation secondaire de leurs données de santé et faciliter l'exercice des droits			
29.	Charger le HDH de mettre en place un portail de transparence centralisé recensant les traitements sur les données de santé.	HDH	Réglementaire
30.	Étudier la mise en place d'une plateforme centralisée pour l'exercice par les patients de leurs droits concernant leurs données de santé, dont leur droit d'opposition, s'agissant de l'ensemble des traitements recourant à leurs données dans le cadre de recherches, d'études et d'évaluations dans le domaine de la santé.	HDH/DNS/ CNAM	
31.	Envisager, à terme, la possibilité d'une réutilisation à des fins de recherche des données saisies dans l'espace numérique de santé, moyennant le consentement exprès de la personne concernée qui le saisirait sur l'application "Mon espace santé".	CNAM/ DNS	Législatif
32.	Mettre en place dans les espaces numériques de santé une information destinée à leurs titulaires sur l'intérêt de la réutilisation des données de santé à des fins de recherche.	CNAM/ DNS	
Développer et consolider certains usages secondaires			
33.	Élaborer à l'échelle du comité stratégique des données de santé un plan d'enrichissement de la base principale du SNDS et en assurer un suivi régulier.	COSTRAT	
34.	Poursuivre les appels à projet pour développer des bases de données en santé sur de nouveaux champs définis par le comité stratégique des données de santé, comme la santé mentale, les pathologies chroniques ou portant sur certaines zones géographiques peu couvertes comme les départements et collectivités d'Outre-Mer.	COSTRAT/ DREES	
35.	Favoriser l'utilisation de données issues de différents champs intersectoriels (en particulier champ environnemental et champ social), notamment pour développer la prévention en santé, en s'appuyant sur les « hubs » sectoriels pour cartographier les données et favoriser leur interopérabilité.	COSTRAT/ HDH/ DINUM	
36.	Favoriser l'accès pour les chercheurs aux bases de données de statistique publique comme l'EDP-Santé ou le répertoire de gestion des carrières uniques (RGCU).	DSS/DREES	Réglementaire (art R. 161-69- 12 CSS)

n°	Recommandation	Autorité responsable	Vecteur juridique (le cas échéant)
37.	Renforcer l'accompagnement de la DREES et du Health data hub auprès des administrations publiques et des agences sanitaires, pour favoriser l'exploitation des données de santé pour le pilotage des politiques publiques.	DREES/HDH	

SOMMAIRE

PRÉAMBULE	3
SYNTHÈSE.....	5
RECOMMANDATIONS DE LA MISSION	13
INTRODUCTION	21
1 LES DONNEES DE SANTE EN FRANCE : UNE RICHESSE INSUFFISAMMENT EXPLOITEE.....	23
1.1 L'UTILISATION SECONDAIRE DES DONNEES DE SANTE : UN POTENTIEL MAJEUR	23
1.1.1 Qu'entend-on par utilisation secondaire des données de santé ?.....	23
1.1.2 Des données de santé aux usages variés et croissants, dans un contexte de développement de l'intelligence artificielle.....	28
1.1.3 Un potentiel macroéconomique important, mais encore mal appréhendé	35
1.2 UN PATRIMOINE DE DONNEES DE SANTE IMPORTANT, DONT L'EXPLOITATION EST NOTAMMENT FREINEE PAR UN DELAI D'ACCES DECOURAGEANT	36
1.2.1 Un paysage foisonnant et peu interopérable des bases de données en santé, récemment enrichi par des entrepôts de données de santé	36
1.2.2 Des bases de données encore sous-exploitées et peu appariées, dont l'interopérabilité n'est pas suffisamment pensée ab initio	49
1.2.3 Un accès aux données de santé qui s'est allongé malgré la création de la plateforme des données de santé (Health Data Hub).....	53
1.3 UN ENVIRONNEMENT INSTITUTIONNEL TROP PEU COOPERATIF ET UN PRINCIPE D'OUVERTURE DES DONNEES ENCORE A ASSEOIR	64
1.3.1 La Plateforme des données de santé (Health Data Hub): de premières réalisations encourageantes	64
1.3.2 Une perspective de centralisation des données qui suscite des crispations importantes, et le maintien de réflexes « propriétaires »	65
1.3.3 Des financements publics importants dans le cadre du plan d'investissement France 2030, mais une mauvaise articulation de certaines initiatives.....	69
1.3.4 Des partenariats économiques tâtonnants, parfois décourageants et peu compétitifs, en l'absence de modèle de tarification lisible.....	73
1.3.5 Un consentement des citoyens complexe.....	76
2 DECLOISONNER LES DONNEES DE SANTE ET FEDERER L'ECOSYSTEME	79
2.1 UN PREALABLE : INSTAURER DAVANTAGE DE CONFIANCE EN ARBITRANT LA QUESTION DE L'HEBERGEMENT DE LA COPIE DE LA BASE PRINCIPALE DU SNDS PAR LE HEALTH DATA HUB (HDH) ..	79
2.2 ACCELERER LES DELAIS DE MISE A DISPOSITION DES DONNEES DE SANTE ET ALLEGER LES PROCESSUS REGLEMENTAIRES	84
2.2.1 Alléger le processus réglementaire et généraliser les procédures simplifiées, pour que l'autorisation devienne l'exception	84
2.2.2 Simplifier le processus d'évaluation scientifique et éthique, dans une logique de subsidiarité	92
2.2.3 Accélérer la mise à disposition des données de la base principale du SNDS	95
2.3 REFONDER ET RENFORCER LE POSITIONNEMENT DU HEALTH DATA HUB EN ASSUMANT UNE LOGIQUE PLUS FEDERATIVE, POUR EN FAIRE UN VERITABLE PILOTE OPERATIONNEL DE L'ECOSYSTEME.....	101
2.3.1 Lever l'obligation de centralisation auprès du Health data hub pour l'extension du catalogue du SNDS.....	101

2.3.2	<i>Renforcer les missions du Health data hub pour en faire un animateur de la mise en réseau des bases de données de santé et un garant du partage</i>	102
2.3.3	<i>Une évolution institutionnelle plus profonde à anticiper, dans la perspective du règlement européen</i>	104
2.4	REUNIR LES CONDITIONS D'UNE UTILISATION SECONDAIRE OPTIMALE DES DONNEES DE SANTE	106
2.4.1	<i>Conforter le pilotage stratégique interministériel</i>	106
2.4.2	<i>Renforcer la visibilité sur les bases de données et leur interopérabilité</i>	108
2.4.3	<i>Consolider le modèle économique et garantir une reconnaissance scientifique pour encourager au partage</i>	113
2.4.4	<i>Renforcer la transparence vis-à-vis des patients sur l'utilisation secondaire de leurs données de santé et faciliter l'exercice des droits</i>	119
2.4.5	<i>Développer et consolider certains usages secondaires</i>	122
	LISTE DES ANNEXES	134
	LISTE DES PERSONNES RENCONTREES	135
	DOCTRINE DE PREVENTION DES CONFLITS D'INTÉRÊTS	144
	SIGLES UTILISÉS	145
	LETTRE DE MISSION	147

INTRODUCTION

Par lettre en date du 31 mai 2023, le ministre de l'Économie, des finances et de la souveraineté industrielle et numérique, la ministre de l'enseignement supérieur et de la recherche et le ministre de la santé et de la prévention ont lancé une mission visant à poser les bases d'une feuille de route française en matière de réutilisation des données de santé. Animée par Jérôme Marchand-Arvier, la mission était également composée de Stéphanie Allassonnière, d'Anne-Sophie Jannot et d'Aymeril Hoang. Emilie Fauchier-Magnan, de l'inspection générale des affaires sociales, et Julien Fradel, du Conseil d'État, étaient rapporteurs de la mission.

Menée sur la période de juin à novembre 2023, la mission s'est notamment appuyée sur :

- des auditions conduites auprès d'une grande diversité d'acteurs (voir liste des personnes rencontrées à la fin du rapport) comprenant des représentants du champ politique et institutionnel (cabinets ministériels, administrations centrales, autorités administratives indépendantes, agences et opérateurs du champ sanitaire et de la sécurité sociale), des chercheurs et organismes de recherche, des partenaires du système de santé (fédérations hospitalières, conférences des directeurs généraux de CHU et de centres hospitaliers, conférence des présidents de commission médicale d'établissement), des entrepôts de données de santé, des patients, des industriels (syndicats professionnels, entreprises spécialisées dans la gestion, la valorisation et/ou l'hébergement des données de santé), des parlementaires ainsi que des personnalités qualifiées ;
- des consultations écrites (détail des répondants dans la liste des personnes rencontrées à la fin du rapport) par la voie :
 - de questionnaires adressés à des représentants de structures et d'établissements de santé, à des instances ordinales, à des associations de patients... ;
 - d'une enquête auprès des agences régionales de santé afin de recenser les initiatives territoriales en matière de structuration d'entrepôts de données de santé ;
 - d'un questionnaire adressé aux agences sanitaires ;
- des échanges avec d'autres missions en cours, notamment celle sur la rénovation de la recherche biomédicale, confiée à Anne-Marie Armanteras et Manuel Tunon de Lara, et celle sur les usages du SNDS conduite par la Cour des comptes ;
- une riche documentation alimentée par des rapports publics⁴, les travaux du Comité stratégique des données de santé, les publications et réflexions produites le Heath Data Hub....

⁴ Dont un rapport du Haut Conseil de la santé publique sur les registres, un rapport de l'inspection générale des affaires sociales sur les cohortes et un rapport de la commission des affaires sociales du Sénat sur les données de santé.

Dans la dernière phase de ses travaux, la mission a également organisé trois ateliers de co-construction, réunissant les représentants des principales parties prenantes⁵, destinés, dans un cadre collaboratif, à établir des constats partagés et faire émerger des points de convergence et des orientations communes sur deux thématiques cruciales en matière d'utilisation secondaire des données, notamment à des fins de recherche et d'innovation :

- le référencement des bases de données existantes et disponibles au partage et les modalités de leur mise en visibilité ;
- la révision des processus d'accès aux données et les actions de simplification et de réduction des délais, en lien avec la mise en œuvre du règlement européen EHDS.

Les liens d'intérêt des membres de la mission sont détaillés à la fin du rapport, ainsi que les procédures mises en place pour prévenir les éventuels conflits d'intérêt.

En dépit des délais contraints dans lesquels ont dû être organisés ces auditions et ateliers de co-construction, les échanges de la mission avec l'ensemble des parties prenantes à la problématique de la réutilisation des données de santé se sont révélés extrêmement riches et ont fait ressortir de très fortes attentes en faveur d'une meilleure mobilisation du potentiel des données de santé en France.

Au terme de ses travaux, la mission constate que si notre pays dispose d'un patrimoine de données de santé particulièrement riche et reconnu au niveau international, dont la lisibilité reste néanmoins perfectible, son exploitation est entravée par de multiples difficultés, dont des conditions, notamment de délai, de mise à disposition sous-optimales, qui s'expliquent en partie par le fait que les extractions et appariements de données issues de la base principale du SNDS demeurent traitées par la CNAM, faute pour la Plateforme des données de santé (désignée Health Data Hub dans la suite du rapport) de pouvoir, à ce stade, héberger et exploiter une copie de cette base sur sa plateforme technologique (partie 1). Elle considère, dans ce contexte, que le rétablissement de la confiance entre les acteurs implique de résoudre sans tarder la question du transfert d'une copie de la base principale du SNDS au Health Data Hub et de conforter ce dernier dans un rôle d'animateur et de fédérateur de l'écosystème. Toujours dans cette logique de confiance, la mission appelle à donner sa pleine portée à la responsabilisation des utilisateurs des données de santé, notamment par la facilitation du recours aux procédures simplifiées d'accès aux données de santé, ainsi qu'au principe de subsidiarité dans l'examen scientifique et éthique des demandes d'accès. Enfin, elle insiste sur le renforcement du pilotage interministériel de notre politique en matière de réutilisation des données de santé pour favoriser leur partage et leur ouverture à d'autres champs (partie 2). Elle formule, à cet effet, 37 recommandations.

Le rapport est complété par des annexes.

⁵ Comité stratégique des données de santé, Health Data Hub, DREES, direction générale de la recherche et de l'innovation, délégation du numérique en santé, CNAM, INSERM, AP-HP, CNIL et France Assos Santé.

1 Les données de santé en France : une richesse insuffisamment exploitée

1.1 L'utilisation secondaire des données de santé : un potentiel majeur

1.1.1 Qu'entend-on par utilisation secondaire des données de santé ?

1.1.1.1 Une acception large de la notion de donnée de santé

Le règlement européen général sur la protection des données (RGPD)⁶ retient une acception large des données de santé définies comme « **les données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de services de soins de santé, qui révèlent des informations sur l'état de santé de cette personne** », que ces données se rapportent à un état de santé passé, présent ou futur⁷.

Souvent qualifiées de « trésor » ou de « mine d'or », les données de santé sont de plus en plus mobilisées pour répondre à une multiplicité de finalités, dont plusieurs alimentent l'espoir d'une santé chaque fois plus performante et mieux adaptée aux besoins de chaque patient. Ces perspectives sont évidentes dans le champ de la recherche et de l'innovation thérapeutique, mais la pandémie de covid-19 a également mis en lumière l'importance des données de santé dans le pilotage des politiques publiques et l'adaptation du système de santé aux crises sanitaires.

Face à ces potentialités nées de l'exploitation des données de santé, hautement génératrices de valeur, nombre des personnalités auditionnées par la mission n'ont pas manqué de rappeler que les données de santé restent un **bien commun**. Le législateur a d'ailleurs pris soin d'interdire toute cession à titre onéreux de données de santé identifiantes⁸.

Sensible par nature, la donnée de santé doit, pour être traitée, faire l'objet, conformément au RGPD, d'opérations destinées à protéger la personne concernée contre toute possibilité de ré-identification. Si l'anonymisation, rendant impossible toute ré-identification, est utilisée notamment pour la production de données agrégées dont la réutilisation n'est soumise à aucune restriction mais dont le niveau de détail a nécessairement été altéré, la pseudonymisation, en

⁶ Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données

⁷ Aux termes du considérant 35 de ce règlement, « cela comprend des informations sur la personne physique collectées lors de l'inscription de cette personne physique en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services (...) au bénéfice de cette personne physique; un numéro, un symbole ou un élément spécifique attribué à une personne physique pour l'identifier de manière unique à des fins de santé; des informations obtenues lors du test ou de l'examen d'une partie du corps ou d'une substance corporelle, y compris à partir de données génétiques et d'échantillons biologiques ; et toute information concernant, par exemple, une maladie, un handicap, un risque de maladie, les antécédents médicaux, un traitement clinique ou l'état physiologique ou biomédical de la personne concernée, indépendamment de sa source, qu'elle provienne par exemple d'un médecin ou d'un autre professionnel de la santé, d'un hôpital, d'un dispositif médical ou d'un test de diagnostic in vitro ».

⁸ Article L. 1111-8 du code de la santé publique, y compris lorsqu'elles permettent d'identifier directement ou indirectement le professionnel prescripteur (article L. 4113-7 du code de la santé publique).

empêchant la ré-identification directe sans informations supplémentaires, permet de préserver le niveau d'information que confère la donnée : toutefois, cette dernière opération, réversible au moins indirectement par croisement des données, n'enlève pas à la donnée de santé son caractère personnel, qui continue donc d'être soumise au RGPD et à la loi « Informatique et libertés ».

**Les méthodes pour protéger les personnes contre une ré-identification
à partir de leurs données de santé selon la CNIL**

1) La pseudonymisation :

La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données à une personne physique identifiée sans information supplémentaire.

En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénoms, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.). La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. Contrairement à l'anonymisation, la pseudonymisation est une opération réversible au moins indirectement : il est possible de retrouver l'identité d'une personne si l'on dispose d'informations supplémentaires.

2) L'anonymisation :

L'anonymisation est un traitement de données personnelles qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute ré-identification de la personne, par quelque moyen que ce soit. Contrairement à la pseudonymisation, l'anonymisation est donc une opération irréversible.

Source : Extraits du site Internet de la CNIL (<https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation>).

1.1.1.2 Des données de santé essentiellement collectées dans le cadre du soin, mais pas seulement

Les données de santé sont aujourd'hui en grande partie produites et collectées à l'occasion du soin, qui constitue le cadre de leur utilisation primaire. À lui seul, le système national d'information inter-régimes de l'assurance maladie (SNIIRAM) regroupe les informations collectées à partir d'1,2 milliard de feuilles de soins⁹.

En matière d'utilisation primaire des données de santé, les pouvoirs publics ont fortement misé sur le numérique pour renforcer le suivi des patients et la coordination de leur parcours de soins. En particulier, la loi n° 2019-774 du 24 juillet 2019 relative à l'organisation et à la transformation du système de santé, dite loi « OTSS », a notamment créé pour tous les assurés sociaux d'un dossier médical partagé (DMP) et d'un espace numérique de santé (ENS) – le premier étant une composante obligatoire du second –. L'ENS, déployé en janvier 2022, a pris le nom de « Mon

⁹ <https://assurance-maladie.ameli.fr/etudes-et-donnees/en-savoir-plus-snds/presentation-systeme-national-donnees-sante-snds>.

espace santé » : il rassemble un ensemble d'outils et services numériques ayant pour objectif de faciliter le partage entre les professionnels suivant un patient des données de santé le concernant.

Les données de santé peuvent néanmoins également être collectées en dehors du soin. C'est en particulier le cas pour les études de type longitudinal telles que les **cohortes**, pour lesquelles des informations relatives à la santé des personnes sont recueillies avec leur consentement exprès, à des fins de recherche ou d'étude statistique, et pour lesquelles le schéma des données collectées est déjà bien connu.

Le projet de règlement européen règlement de la Commission européenne visant à mettre en place l'espace européen des données de santé (« *European Health Data Space* » – EHDS) tient compte de cette réalité, en précisant que les données utilisées à titre secondaire peuvent inclure non seulement des données de santé électroniques à caractère personnel initialement collectées dans le cadre d'une utilisation primaire, mais aussi des données de santé électroniques directement collectées à des fins d'utilisation secondaire.

1.1.1.3 Les tentatives de définition de l'utilisation secondaire des données de santé

S'il n'existe pas de définition consensuelle¹⁰ de l'utilisation secondaire des données de santé dans la littérature scientifique, les travaux sur l'élaboration d'un règlement relatif à la mise en place d'un espace européen des données de santé ont conduit à des définitions de cette utilisation secondaire **par le prisme de ses finalités**.

L'agence du numérique en santé a ainsi défini, dans le cadre de sa consultation menée sur le règlement européen, l'utilisation secondaire ou réutilisation des données de santé comme le « *traitement ultérieur de données de santé électroniques collectées initialement à d'autres fins pour : des statistiques publiques dans le secteur de la santé ; des activités d'intérêt public dans le domaine de la santé, telles que la protection contre les menaces transfrontalières ou la surveillance de la santé publique ; la recherche scientifique ayant trait aux secteurs de la santé ; des activités de développement et d'innovation de produits ou services contribuant à la santé publique ou à la sécurité des soins de santé, des médicaments ou des dispositifs médicaux ou des activités d'éducation dans le secteur de la santé* »¹¹.

C'est également la logique de l'article 34 du projet de règlement EHDS.

¹⁰Voir notamment Regina Becker et al., « Secondary Use of Personal Health Data: When Is It "Further Processing" Under the GDPR, and What Are the Implications for Data Controllers? », *European Journal of Health Law*, 30(2), 129-157 : <https://doi.org/10.1163/15718093-bja10094>.

¹¹ <https://participez.esante.gouv.fr/project/chapitre-4-utilisation-secondaire-des-donnees-de-sante-electroniques/presentation/presentation>.

Les finalités de l'utilisation secondaire des données de santé inscrites à l'article 34 de la proposition de règlement relatif à l'espace européen des données de santé

1. Les organismes responsables de l'accès aux données de santé ne donnent accès aux données de santé électroniques énumérées à l'article 33 que si la finalité prévue du traitement poursuivi par le demandeur est conforme :

- a) aux activités pour des raisons d'intérêt public dans le domaine de la santé publique et de la santé au travail, telles que la protection contre les menaces transfrontières graves pour la santé, la surveillance de la santé publique ou la garantie d'un niveau élevé de qualité et de sécurité des soins de santé et des médicaments ou dispositifs médicaux ;
- b) au fait d'aider les organismes du secteur public ou les institutions, organes et organismes de l'Union, dont les autorités réglementaires, dans le secteur de la santé ou des soins, à accomplir les tâches inscrites dans leur mandat ;
- c) au fait de produire des statistiques officielles à l'échelon national, plurinational et de l'Union en rapport avec les secteurs de la santé ou des soins ;
- d) aux activités d'éducation ou d'enseignement dans les secteurs de la santé ou des soins ;
- e) à la recherche scientifique ayant trait aux secteurs de la santé ou des soins ;
- f) aux activités de développement et d'innovation pour les produits ou services contribuant à la santé publique ou à la sécurité sociale, ou à la garantie d'un niveau élevé de qualité et de sécurité des soins de santé, des médicaments ou des dispositifs médicaux ;
- g) à la formation, au test et à l'évaluation des algorithmes, entre autres dans les dispositifs médicaux, les systèmes d'intelligence artificielle (IA) et les applications de santé numériques, à la contribution à la santé publique ou à la sécurité sociale, ou à la garantie d'un niveau élevé de qualité et de sécurité des soins de santé, des médicaments ou des dispositifs médicaux ;
- h) à la fourniture de soins de santé personnalisés consistant à évaluer, à maintenir ou à rétablir l'état de santé des personnes physiques, sur la base des données de santé d'autres personnes physiques.

Il faut ajouter aux finalités prévues par cet article 34 la production d'indicateurs et l'amélioration du pilotage stratégique de l'activité de soins, qui constitue également une finalité essentielle.

1.1.1.4 Une séparation entre utilisation primaire et utilisation secondaire à réinterroger

Si la distinction entre utilisation primaire et utilisation secondaire est cardinale et fondée, il ne faut pas que cette distinction soit excessivement stricte, car tenir compte du contexte de la collecte des données permet d'améliorer la pertinence de la réutilisation des données, et la réutilisation peut réciproquement améliorer l'usage primaire des données. **Plus l'usage secondaire est en effet proche de son usage primaire, plus il génère de la valeur.**

À titre d'exemple, l'AP-HP mobilise l'exploitation des données de son entrepôt de données de santé au service de l'amélioration de la qualité des soins : des retours réguliers de cette exploitation vers les professionnels de santé permettent ainsi à ces derniers de réinterroger leurs

pratiques, comme l'illustre le développement d'un outil d'aide au repérage des infections du site opératoire mis en place à partir des données de l'entrepôt. En outre, en prenant la mesure des bénéfices qu'ils peuvent tirer dans leur pratique quotidienne de l'exploitation des données de l'entrepôt de la structure à laquelle ils appartiennent, les professionnels de santé sont incités à renforcer la qualité de l'information médicale qu'ils renseignent, en complétant par exemple certaines indications cliniques. Cette « **rétroaction** » de l'utilisation de la donnée de santé doit être favorisée¹².

Plus globalement, le principe que toute utilisation primaire doit pouvoir conduire à une utilisation secondaire doit être posé. C'est la logique de la loi « OTSS »¹³ du 24 juillet 2019 et du projet de règlement européen EHDS qui intègre, à son article 33, dans le champ des données devant être mises à disposition à des fins d'utilisation secondaire les données issues des dossiers médicaux électroniques¹⁴.

Cette perspective de mobilisation à des fins d'usage secondaire de données collectées à titre primaire doit conduire à envisager de mobiliser à terme les **données de santé saisies dans l'espace numérique de santé (ENS) pour une utilisation secondaire, en particulier pour la recherche. Ce n'est actuellement pas possible** (article L. 1111-14 du code de la santé publique (CSP)) et cette restriction se comprend au lancement de l'ENS, afin de favoriser sa montée en puissance et sa bonne appropriation par les Français. Mais l'ENS, si sa montée en puissance se confirme, sera demain une source extrêmement riche, en particulier pour compléter les données du SNIIRAM avec des données médicales. Le cadre d'une telle évolution doit être approfondi – on pourrait notamment imaginer que lors d'ouverture de son espace, chaque personne puisse explicitement consentir à ce que tout ou partie des données qu'il contient puisse être utilisées à des fins secondaires – mais une telle réflexion nous semble devoir être envisagée rapidement, afin de prévoir les évolutions juridiques et techniques nécessaires.

¹² En application de l'article L. 1130-5 du code de la santé publique, créé par la dernière loi de bioéthique (n° 2021-1017 du 2 août 2021 relative à la bioéthique), en cas de découverte, au cours de la réalisation d'une recherche, de caractéristiques génétiques pouvant être responsables d'une affection justifiant des mesures de prévention ou de soins au bénéfice de la personne ou de membres de sa famille potentiellement concernés, il appartient au responsable du programme de recherche exploitant ces données génétiques de contacter le médecin détenteur de l'identité de la personne afin que celui-ci porte à la connaissance de la personne, si elle ne s'y est pas opposée, l'existence d'une information médicale la concernant et l'invite à se rendre à une consultation de génétique.

¹³ En prévoyant, au 6° de l'article L. 1461-1 du code de la santé publique, que le SNDS rassemble et met à disposition notamment « les données destinées aux professionnels et organismes de santé recueillies à l'occasion des activités mentionnées au I de l'article L. 1111-8 du présent code donnant lieu à la prise en charge des frais de santé en matière de maladie ou de maternité mentionnée à l'article L. 160-1 du code de la sécurité sociale et à la prise en charge des prestations mentionnées à l'article L. 431-1 du même code en matière d'accidents du travail et de maladies professionnelles ».

¹⁴ Qui correspondent à l'« ensemble de données de santé électroniques relatives à une personne physique collectées dans le système de santé et traitées à des fins de soins de santé » (point m) du paragraphe 2 de l'article 2 du projet de règlement.

Les données de « Mon espace santé » : des données réservées à un usage primaire

Créé automatiquement pour l'ensemble des assurés sociaux depuis le 1^{er} janvier 2022 sous réserve de l'exercice par ces derniers de leur droit d'opposition, l'ENS, dénommé « Mon espace santé », comporte les données du DMP au sein duquel chaque professionnel de santé est tenu, en application de l'article L. 1111-15 du CSP, de reporter « à l'occasion de chaque acte ou consultation, les éléments diagnostiques et thérapeutiques nécessaires à la coordination des soins de la personne prise en charge ».

À l'heure actuelle, les conditions de partage des données saisies dans le DMP sont strictement définies par le code de la santé publique : en vertu de l'article L. 1111-17, l'accès est réservé au professionnel participant à la prise en charge d'une personne en application des articles L. 1110-4 et L. 1110-12, sous réserve du consentement de la personne concernée préalablement informée, et, aux termes de l'article L. 1111-18, l'accès au dossier médical partagé ne peut être exigé en dehors des cas prévus aux articles L. 1111-15 et L. 1111-16 [c'est-à-dire dans le cadre du parcours de soins de la personne concernée et dans le cadre du suivi de cette personne par son médecin traitant ou le médecin coordonnateur en établissement d'hébergement pour personnes âgées dépendantes], même avec l'accord de la personne concernée.

95 % des citoyens disposent aujourd'hui d'un ENS, 2 % seulement ayant exercé leur droit d'opposition et les 3 % restants correspondant aux assurés sociaux que l'assurance maladie n'a pu contacter, faute de contact disponible. 12 % des titulaires ont consulté au moins une fois leur ENS, et chaque titulaire a la possibilité de verser des données de santé le concernant, notamment des déterminants de santé tels que le poids, la taille, les comportements (consommation d'alcool, de tabac...), certaines données pouvant, le cas échéant, être versées, avec son accord, par l'intermédiaire d'applications de santé référencées pour être connectées à Mon espace santé.

1.1.2 Des données de santé aux usages variés et croissants, dans un contexte de développement de l'intelligence artificielle

Les exemples d'utilisation secondaire des données de santé permettent d'illustrer les **apports variés et croissants de la réutilisation des données de santé**, qui ont été particulièrement mis en lumière pendant la crise sanitaire Covid 19.

Ainsi, des études menées sur les données du système national des données de santé ont permis de démontrer quels étaient les facteurs de risque vis-à-vis du Covid-19 ou encore de prouver l'efficacité des vaccins. L'utilisation des données des établissements de santé a également été essentielle pour le **pilotage de la gestion de crise**, tout comme les remontées des données de biologie, en particulier les données relatives aux tests.

Dans le même champ, la mobilisation des données de santé constitue un levier majeur pour la **surveillance sanitaire et la politique de santé publique**, notamment pour la sécurisation des produits de santé. On peut mentionner l'exemple emblématique du Médiateur, la surmortalité liée au médicament ayant été confirmée par une exploitation des données du SNDS, mais il existe de nombreuses autres illustrations en matière de pharmacovigilance. Ainsi, une étude menée sur plus d'un million de patients a montré que le risque d'infections graves est augmenté de plus de 30%

chez les enfants prenant des inhibiteurs de pompe à protons, un médicament très prescrit chez les enfants, justifiant une prudence dans sa prescription¹⁵.

L'utilisation de données de vie réelle pour l'évaluation des produits de santé est également croissante, même si cela nécessite une évolution des doctrines d'évaluation des autorités, traditionnellement fondées sur l'utilisation de données issues d'essais cliniques randomisés plutôt que sur des données observationnelles. L'évolution des doctrines est progressive, mais, aux Etats-Unis, de premières décisions fondées sur des données de vie réelle ont été récemment enregistrées¹⁶.

La réutilisation des données de santé permet également d'**améliorer la qualité des soins**. Le projet Deep Sarc, porté notamment par le centre Léon Bérard, vise à l'identification des meilleurs schémas thérapeutiques pour le traitement des sarcomes, en croisant une base nationale sur les patients atteints de sarcome (alimentée par des centres experts) et les données de vie réelle du SNDS. L'étude, dont la réalisation est accompagnée par la plateforme de données de santé (par la suite désignée Health data hub), doit permettre d'évaluer si l'administration d'un traitement adjuvant pour les sarcomes non métastatiques est associée à un gain en survie globale.

A cet égard, **l'intelligence artificielle recèle encore un potentiel très important dans le champ de la santé**, avec le développement de nombreux algorithmes et outils numériques pour améliorer la qualité des soins (prévention, dépistage, aide au diagnostic, suivi des patients...).

L'utilisation de l'intelligence artificielle dans la santé a de multiples cas d'application :

- Développer des dispositifs d'aide à la décision : modèles permettant la stratification des patients, l'anticipation de phénomènes (diagnostic, rechute, effets secondaires, etc..), l'analyse d'imagerie (segmentation, comparaison de formes à une population contrôle) etc... pour assister les professionnels de santé dans leur prise de décisions ;
- Développer la médecine de précision : personnaliser les traitements en fonction de l'état exhaustif des connaissances sur les patients, optimiser la fenêtre thérapeutique ;
- Accélérer la recherche et développement de nouveaux outils thérapeutiques qu'ils soient des médicaments (accélération du screening de molécules ou de recherche de cibles) ou de dispositifs médicaux (pacemakers ou stents intelligents, pancréas artificiel pour la stabilisation des patients diabétiques, etc..) ;
- Aider à la formation médicale en proposant des scénarios d'apprentissage interactifs et réalistes (réalité augmentée, apprentissage sur des données artificielles mais reflétant une réalité) ;
- Assister les chirurgiens à la planification pré et per - opératoire ;

¹⁵ <https://pubmed.ncbi.nlm.nih.gov/37578761/>

¹⁶ Selon une étude publiée en janvier 2022 dans la revue *Clinical Pharmacology and Therapeutics*, un total de 116 autorisations de produits de santé par la Food and Drug Administration (FDA) ayant inclus des données de vie réelle auraient été enregistrées sur la période de janvier 2019 à juin 2021 : 88 d'entre elles auraient mobilisé des données de vie réelle dans l'évaluation de l'efficacité et de la sécurité des produits concernés, et, pour 65 de ces autorisations, les études de données de vie réelle auraient pesé sur la décision finale de la FDA. Source : Purpura CA, Garry EM, Honig N, Case A, Rassen JA. The Role of Real-World Evidence in FDA-Approved New Drug and Biologics License Applications. Clin Pharmacol Ther. 2022 Jan;111(1):135-144. doi: 10.1002/cpt.2474. Epub 2021 Nov 22. PMID: 34726771; PMCID: PMC9299054.

- Introduire une méthodologie d'essais cliniques du futur pour évaluer l'efficacité et la sécurité de nouvelles molécules ou dispositifs médicaux en utilisant des cohortes de patients partiellement ou complètement artificiels.

Dans ces applications, l'utilisation de données de santé est fondamentale. Chacun de ces objectifs ne peut être atteint sans une confrontation à la donnée que ce soit pour la phase d'apprentissage, pour calibrer les modèles, ou pour leur validation, pour évaluation des performances sur des données réelles.

A titre d'exemples, la société Therapixel a développé des algorithmes pour appuyer les radiologues pour le dépistage du cancer du sein, en s'appuyant sur la collecte de grandes bases de données d'imagerie. La société Implicit a quant à elle développé des algorithmes à partir de données de santé (notamment les données du SNDS) pour améliorer la télésurveillance de patients avec des stimulateurs cardiaques implantables.

Par ailleurs, **de nouveaux usages de l'intelligence artificielle aux données de santé se développent pour les essais cliniques ou pour la création de jumeaux numériques.**

L'utilisation de l'intelligence artificielle pour les essais cliniques du futur

Dans un contexte de coûts croissants des essais cliniques liés à la difficulté de recrutement des patients, l'intelligence artificielle pourrait permettre de nouveaux designs de protocoles d'essais cliniques en particulier intégrant des bras contrôles partiellement ou intégralement virtuels et permettant une inclusion optimisée à la fois en termes de profil patient mais également de cofacteurs.

L'augmentation numérique du nombre de patients devrait permettre de répondre à la problématique de recrutements insuffisants, conduisant à des tailles d'échantillon faibles et rendant les analyses statistiques classiques peu puissantes. L'augmentation des données est aussi une solution pour l'entraînement et la validation des dispositifs médicaux pour leur déploiement et leur mise sur le marché, car ces modèles doivent généralement être entraînés sur d'énormes ensembles de données pour éviter un sur-apprentissage ou pour donner des résultats statistiquement significatifs¹⁷.

A côté des données collectées, il est possible de générer numériquement des données dites artificielles : ces données artificielles sont des informations ou des ensembles de données créés numériquement pour simuler des caractéristiques et des structures similaires à celles des données réelles.

¹⁷ C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

Dans une perspective d'essais cliniques du futur, ces techniques pourraient être utilisées pour réduire le nombre de personnes recrutées¹⁸. Par exemple, si dans une phase 3, il est nécessaire d'avoir 600 patients, 300 dans chaque bras, on pourrait envisager de ne recruter que 100 patients du bras contrôle tout en gardant les 300 patients dans le bras traité. Le bras traité garderait sa diversité permettant de suivre l'efficacité ainsi que les possibles effets indésirables de la nouvelle molécule. En revanche, en recrutant 100 patients représentatifs de la population d'intérêt dans le bras contrôle et en augmentant artificiellement cette population jusqu'à 300 individus, le recrutement est réduit d'un tiers. Ce gain de recrutement permet à la fois un gain de temps et un gain financier permettant d'envisager une mise sur le marché plus rapide de molécules et de dispositifs médicaux et de développer davantage les essais cliniques pour augmenter les potentielles thérapies.

La création de jumeaux numériques

Les progrès rapides de la simulation numérique ouvrent la voie à la possibilité de créer des jumeaux numériques de patients ou d'organes.

De nombreux jumeaux simulés sont adaptés pour prédire des propriétés d'un organe et préparer les décisions médicales. Les simulations cardiaques sont une réalisation majeure dans ce domaine. Les modèles sont calibrés en comparant aux données réelles de l'organe.

Les jumeaux simulés peuvent aussi donner lieu à des applications chirurgicales portant sur divers organes tels que le foie ou le cerveau.

Des applications majeures sont aussi envisagées dans divers domaines, donnant lieu à de nombreux projets de recherche : la compréhension de la variabilité au sein d'une population saine ou malade ^{19 20}, l'anticipation de diagnostics (par exemple pour les maladies neurodégénératives [docteurmemo.fr](https://www.docteurmemo.fr)²¹), l'utilisation de ces jumeaux numériques comme bras placebo optimal dans les essais cliniques.

Enfin, les données de santé devraient être utilisées de manière croissante par les autorités publiques dans un contexte de développement de financements à la qualité des soins et de nouvelles modalités de financement des produits de santé (cf partie 2.4.5.3). Le développement de l'intelligence artificielle générative fait également l'objet d'une attention particulière par la

¹⁸ Dimitris Papadopoulos, and Vangelis D. Karalis. Variational Autoencoders for Data Augmentation in Clinical Studies. MDPI Applied sciences

¹⁹ Bishesh Khanal, Marco Lorenzi, Nicholas Ayache, Xavier Pennec. A biophysical model of brain deformation to simulate and analyze longitudinal MRIs of patients with Alzheimer's disease. NeuroImage, 134, 35-52, 2017

²⁰ V Debavelaere, S Durrleman, S Allassonnière. Learning the clustering of longitudinal shape data sets into a mixture of independent or branching trajectories. International Journal of Computer Vision 128, 2794-2809

²¹ Etienne Maheux, Igor Koval, Juliette Ortholand, Colin Birkenbihl, Damiano Archetti, Vincent Bouteloup, Stéphane Epelbaum, Carole Dufouil, Martin Hofmann-Apitius & Stanley Durrleman. Forecasting individual progression trajectories in Alzheimer's disease. Nature Communications volume 14, Article number: 761 (2023)

puissance publique, avec l'inauguration d'un comité de l'intelligence artificielle générative en septembre 2023.

Le tableau qui suit résume les finalités très variées de l'utilisation secondaire des données de santé illustrées précédemment par quelques exemples.

Tableau 1 : Typologie des utilisations secondaires des données de santé, par domaine d'application

Axe	Domaine	Exemples
Recherche et développement d'outils et produits innovants	Recherche non interventionnelle- notamment à travers la constitution de cohortes et de registres	Etude des profils des patients pour une pathologie donnée, Evaluation des liens entre l'exposome des individus et leur état de santé Analyse des parcours d'errance diagnostique et de prise en charge ainsi que du fardeau d'une maladie
	Recherche clinique	Repérage de patients en vue de leur inclusion dans des essais cliniques (« pré-screening » des dossiers médicaux hospitaliers pour faciliter l'enrôlement des patients Utilisation des données rétrospectives d'essais passés (bras synthétiques) permettant l'accélération de la mise sur le marché Simulation de « bras virtuels », dans le cadre d'études dites « in silico » comprenant des patients fictifs générés à partir des distributions des données observées
	Recherche translationnelle	Recherche de biomarqueurs, de cible thérapeutique
	Recherche appliquée et développement d'innovations	Mise au point et entraînement d'algorithmes pour le développement de modèles prédictifs et de solutions d'intelligence artificielle, par ex pour <ul style="list-style-type: none"> ○ Mieux détecter les symptômes d'une pathologie ; ○ Faire un suivi prédictif de l'évolution d'une maladie ; ○ Aider à la décision clinique ; ○ Formuler des propositions thérapeutiques plus personnalisées ; ○ Optimiser une prise en charge à l'aide un jumeau numérique

Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé

Evaluation en vie réelle des produits de santé	Evaluation des produits de santé	Mesure de l'efficacité d'un médicament ou d'un dispositif médical en conditions réelles, sur une large population Suivi de produits de santé, notamment les molécules et thérapies innovantes, dans le cadre des accès précoces et du paiement à la performance
	Pharmacovigilance/ sécurité sanitaire	Surveillance de la sécurité des produits de santé mis sur le marché : détection et analyse des effets indésirables (exemple : Mediator), surveillance des mésusages, etc.
Pilotage du système de soins et amélioration de la qualité des soins	Santé publique et épidémiologie	Description de l'histoire naturelle des maladies Analyse de facteurs de risques populationnels Renforcement des capacités d'anticipation et de pilotage des crises sanitaires et épidémiques
	Pertinence et qualité des soins	Amélioration de la connaissance des parcours et identification des prises en charge sous-optimales Analyse des pratiques médicales Recueil d'indicateurs renseignant sur la qualité des soins
	Pilotage de l'offre de soins	Production d'indicateurs de pilotage et d'outils de suivi (ex : sur les urgences ou les soins non programmés, sur les disparités territoriales)
	Prévention	Meilleur ciblage des politiques de prévention en santé Analyses territoriales
	Formation des professionnels de santé	Proposition de scénarios d'apprentissage interactifs et réalistes

Source : Mission

1.1.3 Un potentiel macroéconomique important, mais encore mal appréhendé

Il existe un consensus sur le fait que le marché mondial de la donnée est en croissance forte et offre des perspectives croissantes, même s'il n'existe pas d'évaluation fiable de sa valeur. Les estimations disponibles attestent **d'écarts très importants selon les évaluations**. Elles sont principalement centrées sur l'utilisation des données de vie réelle par l'industrie de santé, qui ne constitue qu'un des utilisateurs de données de santé, comme on l'a vu dans la partie précédente.

Ainsi, l'entreprise Markets and Markets²² estime le marché mondial des services liés à l'utilisation des données de vie réelle à 1,5 Md \$ en 2022, avec une prévision 2,9 Md\$ en 2027. Selon l'analyse de Fortune Business insights²³, l'évaluation de ce marché est de 15 Mds \$ en 2022 (dont 6,6 Mds \$ en Amérique du Nord), avec une prévision de 36 Mds \$ en 2030. L'importance de ces écarts s'explique par une différence dans les champs retenus, à la fois dans le champ des données (bases de données seules dans le premier cas, ou bases et études observationnelles dans le second cas) et dans le périmètre des services (services directement liés à l'exploitation des données, dans le premier cas, ou services plus larges comme l'aide à la définition de la stratégie d'accès au marché, dans le second cas).

Ces analyses s'accordent toutefois sur le fait que la crise Covid a constitué un accélérateur important pour la réutilisation des données de santé dans le champ de l'industrie pharmaceutique, du fait de l'augmentation du nombre d'essais cliniques et de la croissance de demande de nouveaux médicaments et de nouveaux vaccins, qui se traduisent par une augmentation des dépenses de recherche et de développement.

Le potentiel économique de l'utilisation secondaire des données de santé est également attesté par le développement dynamique d'entreprises mobilisant les données de santé avec un usage secondaire, et par leur valorisation économique ou leur capacité à lever des fonds (détail ci-dessous).

Le développement d'entreprises avec un usage secondaire des données de santé

En 2022, les entreprises les mieux valorisées étaient américaines, les six licornes (start-ups valorisées à plus de 1 milliard de dollars) étant situées aux Etats-Unis. Les entreprises européennes ayant réussi des levées de fonds de plus de 100 M€ étaient britanniques. En France, une des entreprises les plus en pointe sur le sujet est l'entreprise Cardiologs, avec 23 M€ levés.

Beaucoup de ces entreprises sont des plateformes, avec différents modèles :

- mise à disposition de données structurées et anonymisées ainsi que des outils d'intelligence artificielle entraînés sur ces données ;
- structuration des données des clients et restitution, avec mise à disposition d'outils.

²² RWE Solutions Market Global Forecast for 2026

²³ « Real world evidence solutions market size, share & covid-19 impact analysis, by type and regional forecast, 2023-2030 », Fortune business insights, 2023

Les données le plus souvent utilisées par ces entreprises sont les dossiers médicaux électroniques, même si les données génomiques semblent prendre de plus en plus d'importance.

Sur le champ de l'industrie pharmaceutique, les startups identifiées comme étant les plus valorisées se situent sur le champ de la découverte de nouveaux médicaments. Par exemple, la société hongkongaise InSilico Medecine a ainsi développé grâce à de l'intelligence artificielle un médicament candidat en moins de 18 mois pour un coût de 2,6 M \$ ce qui est environ 3 fois moins long et plus de 100 fois moins onéreux qu'un développement classique.

Sur le champ des dispositifs médicaux utilisant de l'intelligence artificielle, la radiologie est la spécialité médicale qui a récolté le plus de certifications aux Etats-Unis et de marquages CE ces dernières années sur ces sujets.

Source : Extraits de la note "Quelle valorisation pour l'utilisation secondaire des données de santé ? », Veltys, 2022

1.2 Un patrimoine de données de santé important, dont l'exploitation est notamment freinée par un délai d'accès décourageant

1.2.1 Un paysage foisonnant et peu interopérable des bases de données en santé, récemment enrichi par des entrepôts de données de santé

Un bref panorama des données de santé est présenté ci-dessous. Il n'a pas vocation à être exhaustif, dans un contexte de foisonnement d'initiatives, mais vise une présentation synthétique du paysage, avec les principaux atouts et faiblesses des bases françaises. Conformément à la lettre de mission, une analyse plus approfondie des entrepôts de données de santé, de création plus récente, a été réalisée par la mission. Le détail de l'analyse est présenté en annexe 1.

1.2.1.1 La base principale du système national des données de santé (SNDS) est d'une très grande richesse malgré une faible médicalisation

Le système national des données de santé : base principale, SNDS élargi et catalogue du SNDS

Le SNDS a été créé par la loi du 26 janvier 2016²⁴, en rassemblant sous une même architecture plusieurs fichiers existants ou en cours de constitution²⁵ :

- Le système national d'information interrégimes de l'assurance maladie (Sniiram), qui contient les données relatives à toutes les dépenses de l'assurance maladie et les données du programme de médicalisation des systèmes d'information (PMSI) d'analyse de l'activité des établissements de santé ;
- Les données relatives aux causes de décès du CépiDc (Centre d'épidémiologie sur les causes médicales de décès de l'Inserm) ;
- à compter de fin 2024-début 2025, les données médico-sociales liées au handicap fournies par les maisons départementales des personnes handicapées (MDPH).

Cette architecture est aujourd'hui désignée comme « SNDS historique » ou « base principale du SNDS » qui intègre également, sans que cet ajout n'ait été codifié²⁶, des traitements de données mis en œuvre durant la crise sanitaire (SI-Dep, Vaccin-covid notamment).

Depuis 2019, la caisse nationale d'assurance maladie (CNAM) et la plateforme des données de santé (dite Health data hub) sont responsables conjoints de traitement du SNDS.

D'une profondeur historique de 20 ans, les données constituant la base principale du SNDS sont pseudonymisées par la CNAM (aucun nom, prénom, adresse ni numéro de sécurité sociale n'y figure), mais elles restent indirectement identifiantes²⁷. C'est pourquoi leur accès est encadré : hormis certains organismes publics (listés dans un décret) qui disposent d'accès permanents aux données du SNDS, l'accès à ces données suppose une autorisation CNIL ou, pour certains usages bien définis, le respect de méthodologies de référence définies par la CNIL. En cas d'autorisation de la CNIL, un comité éthique et scientifique formule également un avis sur la méthodologie de l'étude et sur la finalité d'intérêt public.

²⁴ Loi n° 2016-41 de modernisation de notre système de santé

²⁵ Selon les termes de l'article L 1471-1 du code de la santé publique, le SNDS intègre également un échantillon représentatif des données de remboursement par bénéficiaire transmises par des organismes d'assurance maladie complémentaire et défini en concertation avec leurs représentants.

²⁶ L'arrêté du 12 mai 2022 relatif aux données alimentant la base principale et aux bases de données du catalogue du système national des données de santé précise certes que les données de Vaccin-covid et SI-Dep alimentent la base principale, mais le code de la santé publique n'a pas été modifié.

²⁷ Du fait de la richesse des données contenues dans le SNDS historique, il est possible d'identifier des personnes connues par ailleurs en croisant certaines informations qui y figurent. Par exemple, dès lors que le système rassemble tous les épisodes de soins d'une personne, si un tiers ayant accès à la base dispose d'une connaissance, acquise par ailleurs, de certains de ces épisodes et si ceux-ci sont suffisamment discriminants, il est possible de repérer la personne concernée dans l'ensemble des enregistrements et d'accéder ainsi à des informations supplémentaires sur sa santé (diagnostics, médicaments pris) au mépris du respect dû à la vie privée. Ce risque est d'autant plus grand que les informations rassemblées sous l'identifiant commun sont nombreuses, et que les données se rapportant une même personne s'accroît. A l'inverse, le risque diminue ou disparaît lors les données sont agrégées. Source : « Rapport sur la gouvernance et l'utilisation des données de santé », Pierre-Louis Bras, 2013

Un échantillon rassemblant les données de la base principale du SNDS pour 2% de la population (Echantillon du système national des données de santé- ESND) est également accessible de manière simplifiée et plus rapide, sous certaines conditions.

La loi du 24 juillet 2019²⁸ a élargi la notion de SNDS, en ajoutant six types de données complémentaires :

- Les données recueillies à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi médico-social ;
- Les données relatives à la perte d'autonomie (si elles sont appariées avec les autres données du SNDS) ;
- Les enquêtes dans le domaine de la santé reliées à des données de la base historique (si elles sont appariées avec les autres données du SNDS) ;
- Des données recueillies lors de dépistages ;
- Les données recueillies par les services de protection maternelle et infantile ;
- Les données issues des dossiers médicaux en santé au travail.

Ainsi, le cadre réglementaire (article R 1461-2 du CSP) distingue désormais la base principale du SNDS, couvrant l'ensemble de la population, et le catalogue du SNDS, ensemble de bases de données ne couvrant pas l'ensemble de la population. A ce stade, le catalogue du SNDS regroupe 12 bases de données (registres, cohortes, enquêtes, données de gestion du système de santé²⁹), listées par arrêté, dont l'intégration à la plateforme du Health data hub est prévue, avec l'objectif d'un appariement avec le SNDS principal. Une seconde liste de bases de données destinées à alimenter la plateforme est en cours de constitution.

En dépit de cette clarification théorique, les acteurs confondent SNDS, base principale du SNDS, catalogue du SNDS et plateforme du Health data hub, de sorte qu'il serait opportun de clarifier ces différentes notions.

Enfin, l'assurance maladie ainsi que l'ATIH mettent en ligne un nombre croissant de jeux de données agrégés issues de ses systèmes d'informations (données sur les dépenses d'assurance maladie, sur les prescriptions de médicaments délivrés en ville, données de biologie, données sur les dispositifs médicaux inscrits à la liste des produits et des prestations, etc.), avec des outils de visualisation des données.

L'ensemble des acteurs souligne que la base principale du SNDS constitue un atout très fort pour la France. Malgré des limites inhérentes au caractère médico-administratif des données (voir infra), elle présente des avantages très importants qui en font une base de données presque sans équivalent dans le monde :

- exhaustivité de la population, des données relatives aux hospitalisations et des traitements médicamenteux prescrits,
- suivi sur une longue période sans interruption, chronologie précise,
- qualité d'ensemble satisfaisante des données et homogénéité globale du codage.

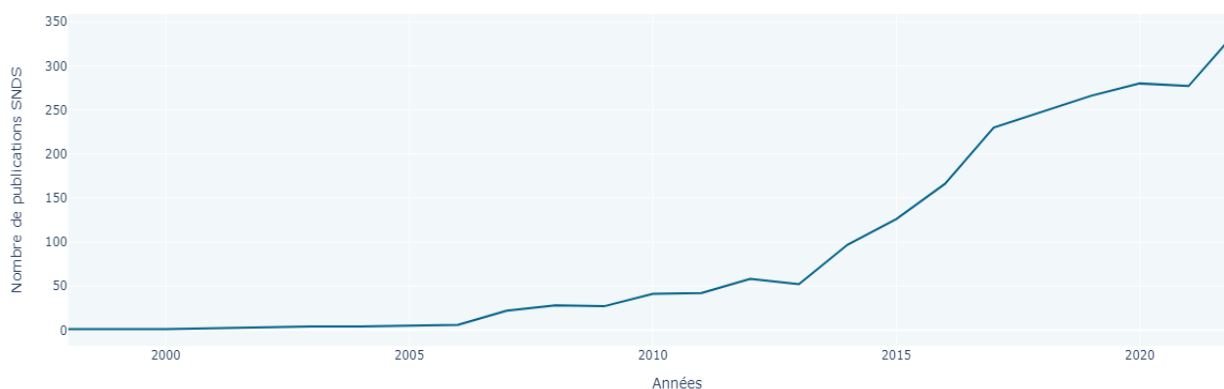
²⁸ Loi n°2019-774 relative à l'organisation et à la transformation du système de santé

²⁹ Le détail des bases est présenté partie 1.3.2.1.

La richesse des données de la base principale du SNDS se traduit par une croissance soutenue de son utilisation et de sa sollicitation à des fins de recherche, comme l'illustre l'augmentation très forte du nombre de publications intégrant des données du SNDS depuis dix ans³⁰. Celle-ci est notamment le fait du groupement d'intérêt scientifique EPI-PHARE, rassemblant depuis 2018 les compétences de l'agence nationale de sécurité du médicament et des produits de santé (ANSM) et de l'assurance maladie, qui publie de nombreuses études à partir de l'exploitation des données du SNDS. A titre d'exemple, une étude portant sur l'ensemble de la population du SNDS réalisée pendant la crise Covid a montré que l'âge était le principal facteur de risque d'hospitalisation ou de décès pour COVID-19³¹.

L'augmentation importante du nombre d'accès permanents au SNDS depuis 2021 (environ 2 500 personnes ayant un accès permanent au SNDS sont recensées par la caisse nationale d'assurance maladie (CNAM)), particulièrement appréciée par les acteurs concernés, contribue aussi à l'augmentation de l'utilisation des données du SNDS.

Graphique 1 : Nombre de publications sur les données du SNDS de 1998 à 2022



Source : Health data hub (données en ligne sur les publications en provenance de PubMed, Lissa et HAL)

Une des principales limites de la base principale du SNDS tient à la faiblesse des données médicales contenues. L'absence d'information sur les diagnostics en médecine de ville, de description des consultations médicales ou encore l'absence des résultats d'examens prescrits constituent des limites importantes. C'est pourquoi les appariements de bases de données médicales avec le SNDS, qui restent aujourd'hui limités et complexes (voir *infra*) constituent une piste particulièrement prometteuse.

L'absence de données relatives aux principaux déterminants de santé des individus (revenus, informations sur les habitudes de vie, comme la consommation de tabac ou d'alcool, l'indice de masse corporelle) limite aussi les analyses. Même si elle a été augmentée, la profondeur historique de 20 ans au maximum pourrait aussi être réinterrogée, notamment du fait de leur disponibilité à

³⁰ Le principal journal de publications est la revue d'épidémiologie et de santé publique, mais le panel intègre des journaux internationaux (Archives of cardiovascular diseases, Fundamental & clinical pharmacology, ...). Les 5 structures publiant le plus d'articles sont l'Inserm, l'assurance maladie (GIP Epiphare), Santé publique France, université de Bordeaux, université Paris cité.

³¹ "Chronic diseases, health conditions and risk of COVID-19-related hospitalization and in-hospital mortality during the first wave of the epidemic in France A cohort study of 66 million people", Lancet regional health Eur, 2021

venir et dans un contexte de croissance des analyses relatives à l'impact des expositions environnementales et à leur retentissement à long terme. De même, la fréquence d'actualisation de certaines données apparaît insuffisante à certains acteurs (données de décès ou données du PMSI, disponibles de manière fiables et complètes annuellement), notamment en période de crise.

Enfin, la base de données est organisée de manière très complexe, ce qui nécessite une expertise et un temps d'appropriation élevé, et limite aussi les capacités d'exploitation de la base.

Mais le principal obstacle à une plus grande réutilisation des données du SNDS tient aujourd'hui au délai d'accès très long aux données (voir *infra* partie 1.2.3).

1.2.1.2 De nombreux registres et cohortes, avec une richesse et une qualité de données très variables, souffrant d'un relatif isolement

Cohortes et registres : définitions et caractéristiques

Une cohorte est un groupe de sujets qui sont suivis individuellement, de manière longitudinale (c'est-à-dire de manière répétée dans le temps), **selon un protocole de recherche préétabli**. Une distinction majeure doit être faite entre cohortes de patients souffrant d'une pathologie particulière (par exemple, la cohorte Memento suit 2 300 patients ayant des signes cliniques pouvant évoquer une forme débutante de la maladie d'Alzheimer ou une maladie apparentée), et cohortes dont les participants sont issus de la population générale (la plus importante en France est la cohorte Constances qui suit 200 000 sujets âgés de 18 à 69 ans consultant les centres d'exams de la sécurité sociale). Comme le rappelait le rapport IGAS/IGESR consacré aux cohortes³², elles ont des objectifs et des caractéristiques très différents : cohortes pluridimensionnelles en population *versus* cohortes spécialisées de personnes atteintes et/ou traitées au titre d'une ou plusieurs pathologies ; cohortes « d'amont » cherchant à évaluer les effets des différents facteurs ou expositions ou cohortes « d'aval » centrées sur l'impact de thérapeutiques ou de politiques publiques ; cohortes fermées, encore ouvertes ou aspirant à l'être à nouveau, etc.

³² « Les cohortes pour la recherche et les études en santé », IGAS-IGESR, 2019

Un registre est un recueil continu et exhaustif de données nominatives, intéressant un ou plusieurs événements de santé dans une population définie, le plus souvent selon un critère géographique, à des fins de recherche et de santé publique. Les registres populationnels ont initialement été conçus dans une perspective de calcul des incidences, prévalences et survie des pathologies recueillies, et afin de répondre à des questions de recherche, grâce à un codage précis des diagnostics. A titre d'exemples, le registre REIN (Réseau Epidémiologie et Information en Néphrologie) géré par l'agence de biomédecine assure le suivi à l'échelle nationale de l'ensemble des patients en insuffisance rénale chronique terminale traités par dialyse ou par greffe, quand le registre des cancers du Tarn vise un recensement exhaustif des cas de cancers dans le département. Les registres populationnels se distinguent des registres des pratiques, qui sont des collectes de données structurées sur la pratique d'une spécialité médicale ou relative à un acte ou un dispositif médical, alimentés par les professionnels de santé, souvent portés par des sociétés savantes et des conseils nationaux professionnels. Ainsi, le conseil national professionnel (CNP) de radiologie et d'imagerie médicale porte 38 registres. Quant à la Fédération des spécialités médicales, sa plateforme héberge 16 registres de pratiques médicales, 4 autres sont en cours de finalisation et trois nouveaux registres sont en cours de développement.

A la différence des cohortes, portées par des équipes de recherches, les registres sont en général portés par des acteurs cliniques, sans adossement systématique à une structure de recherche. D'après un rapport du haut conseil à la santé publique, la cancérologie et les maladies rares sont les deux spécialités médicales les mieux couvertes par les registres.

Les cohortes et les registres sont nombreux en France : le portail France Epidémiologie recensait en octobre 2023 environ 330 cohortes et 110 registres³³. **Ils sont toutefois particulièrement hétérogènes.** La taille des populations suivies est extrêmement variable, de même que leurs objectifs et leurs caractéristiques (cf encadré ci-dessus).

Parmi les atouts de ces bases de données figure **la richesse et la précision des données médicales** contenues, même si elles sont en général ciblées.

Les registres et les cohortes constituent des contributeurs importants en termes de recherche scientifique. Ainsi, le registre des cancers du Tarn a fait l'objet 140 publications recensées sur Medline entre 1998 et 2020. Leurs données sont utilisées par des partenaires locaux (ARS, CPAM, CREAI-ORS Occitanie, réseau régional...) mais aussi de réseaux nationaux ou européens (European Network of cancer registries). La cohorte Constances fait l'objet d'un nombre de publications croissant (près de 40 publications en 2022 contre 25 en 2020 et 2021 ; d'après les données de pubmed) ; elle fait l'objet de plus de 130 collaborations françaises et internationales³⁴, elle est membre de 8 consortiums européens, et ses données sont utilisées par les acteurs publics comme l'ANSES, Santé publique France ou la Haute Autorité de santé.

De manière plus générale, le rapport IGESR/IGAS précité constatait, pour les cohortes financées par le plan Investissement d'avenir, « que la qualité de la production scientifique, la structuration des communautés de recherche autour d'un champ et d'un instrument fédérateurs, ainsi que le développement des partenariats internationaux constitu[aient], dix ans après, trois acquis

³³ Ce recensement n'est pas exhaustif : en 2021, le haut conseil de la santé publique avait identifié 53 registres en complément des 64 registres inclus dans le bilan de l'existant du portail Epidémiologie France.

³⁴ Donnée 2022

appréciables, communs à la plupart des projets. Les cohortes contribuent ainsi à conforter la place de la France dans la recherche médicale, dans des domaines où sa spécialisation est reconnue ».

Registres et cohortes **sont aussi utilisés de manière croissante pour le pilotage des politiques publiques**, du fait de leur maturité croissante, de financements publics plus importants pour les cohortes³⁵, et d'un intérêt croissant pour la mesure des expositions. Ainsi, les données des registres du cancer (31 registres, dont 18 registres généraux) sont utilisées par l'Institut national du cancer et Santé publique France pour l'estimation de la prévalence des cancers.

Les cohortes et registres sont toutefois confrontés à plusieurs enjeux importants.

En premier lieu, beaucoup de cohortes et de registres résultent d'initiatives individuelles de praticiens ou d'équipes isolées, avec, pour les registres, une structuration et une coordination d'ensemble qui reste faible, et une qualité très hétérogène, globalement peu évaluée.

Un des principaux enjeux pour ces bases de données est d'ordre réglementaire. Un nombre important de ces bases n'est pas en adéquation avec le règlement européen pour la protection des données (RGPD) ou avec la réglementation CNIL, notamment s'agissant des registres³⁶. A titre d'illustration, peu de registres peuvent intégrer le catalogue du Health data hub, car ils contiennent des données ne respectant pas les standards de pseudonymisation. De fait, les professionnels (chercheurs ou médecins) porteurs des registres et cohortes sont très en attente d'un appui juridique pour les éclairer sur la réglementation (existence de méthodologies de référence...) et pour les accompagner dans leurs démarches réglementaires.

Le croisement avec d'autres bases de données pour enrichir leur contenu, et le renforcement de leur interopérabilité constituent un enjeu supplémentaire. Le rapport précité sur les cohortes encourageait à développer les croisements de données entre cohortes et d'autres bases de données médico-administratives (SNDS, bases fiscales et sociales de l'INSEE), pour renforcer leur robustesse (augmenter la durée de suivi et les informations disponibles pour chaque individu) et remédier à leur relatif isolement dans le paysage de la statistique nationale ou dans celui de la recherche publique. De tels appariements gagnent à être intégré dès la création du registre (à l'instar de la cohorte Constances, appariée dès l'origine au SNDS), car ils nécessitent le consentement des participants ou patients.

Le chaînage avec des données de vie réelle, notamment à travers les entrepôts de données de santé (voir partie suivante) constitue un enjeu important pour les cohortes et registres, qui pourront à l'avenir s'appuyer sur ces données pour une alimentation plus automatisée de données. Ceci accroît la nécessité d'une interopérabilité et d'une interconnexion entre ces bases de données, et l'utilisation de nomenclatures communes. Or seuls 1/3 des registres examinés par le Haut Conseil à la santé publique en 2021 avaient recours à des nomenclatures pour le codage de leurs variables.

³⁵ A titre d'exemple, le programme d'investissement d'avenir a mobilisé 124 millions d'euros pour le financement de cohortes sur la période 2011- 2018, à travers trois appels à projets distincts (Cohortes, Equipex, Infrastructures nationales Biologie santé).

³⁶ Les registres sont considérés par la CNIL comme des entrepôts de données de santé, quand les cohortes sont qualifiées en général de projets de recherche.

Enfin, **la réutilisation des données des cohortes et des registres reste encore limitée, avec une politique d'ouverture de l'accès à ces données** encore insuffisante (voir partie 1.3.2). Ainsi, d'après l'analyse du haut conseil à la santé publique, en 2021, moins de 20% des registres mettaient leurs résultats en accès libre.

Le projet France cohortes (détail dans l'encadré *infra*, partie 1.2.1.4) constitue une initiative bienvenue de structuration pour répondre à ces différents enjeux, même si son articulation avec le Health data hub mérite d'être clarifiée (cf. partie 1.3.3).

D'autres pays, qui ont historiquement investi dans les cohortes et les registres comme outil de recherche et de santé publique, ont mis en place une structuration nationale favorisant le chaînage avec d'autres bases de données ainsi une politique active d'ouverture de leurs données.

Exemples étrangers relatifs aux cohortes et aux registres

S'agissant des cohortes, on peut notamment citer le cas de la Grande-Bretagne, avec un investissement ancien pour la constitution de grandes cohortes en population, et leur intégration dans une politique de santé publique historiquement beaucoup plus développée. La cohorte UK Biobank, qui regroupe les données de 500 000 volontaires depuis 2004, y compris des données génétiques, biologiques et d'imagerie, et fait l'objet d'une politique active d'ouverture des données, constitue un exemple emblématique. Elle fait l'objet d'un nombre de publication élevé, augmentant de manière exponentielle : 162 en 2017, 313 en 2018, 1368 en 2021, 1788 en 2022. La comparaison avec les cohortes françaises comporte des limites importantes. La cohorte UK Biobank a fait l'objet d'un investissement public largement supérieur à celui des cohortes françaises (évalué à 580 M€ par l'INSERM, contre 55 M€ pour Constances) et de partenariats très développés. Son ancienneté et la mise à disposition de données génétiques leur confèrent aussi une richesse de données beaucoup plus importante. Par ailleurs, le Royaume-Uni offre un environnement scientifique et réglementaire particulièrement favorable aux cohortes, avec la possibilité d'appariements quasi systématiques à des bases de données administratives très riches en informations sociales comme médicales, en provenance notamment du National Health System (NHS).

S'agissant des registres, en Suède et en Norvège, les registres de santé sont gérés par des institutions publiques nationales et régulés par la loi³⁷. Ils sont en nombre réduit (à hauteur d'une dizaine), collectent peu de données mais bénéficient d'une couverture quasi-exhaustive (98% de la population, avec des populations respectivement de 10 et 5 millions d'habitants). En France, les deux registres les plus anciens, pour le cancer et les maladies congénitales, ne couvrent aujourd'hui qu'environ 22% de la population française. Ces registres nordiques ont permis de mettre en évidence l'association entre certaines expositions et événement de santé négatifs pour la population, grâce à un chaînage avec d'autres bases de données. Ainsi, la Suède a établi très tôt un lien entre le tabagisme au cours de la grossesse et le développement de cancers chez les enfants. Ce chaînage leur permet aussi de disposer de registres de cancers intégrant des données sur l'activité professionnelle. L'accès aux données des registres et aux analyses réalisées grâce à ces données est facilité (accès aux données agrégées par le portail Statistics Sweden ; accès aux données individuelles sur demande auprès du responsable du registre). Dans ces pays, les données cliniques nominatives des patients sont recueillies sans leur consentement exprès, ce qui constitue un enjeu par rapport à la réglementation de protection des données.

Source : *Mission à partir des rapports « Les cohortes pour la recherche et les études en santé », IGAS-IGESR, 2019 et « Registres et données de santé : utilité et perspectives en santé publique », Haut comité à la santé publique, 2021*

³⁷ Les registres suédois faisaient l'objet d'un financement de 45 millions de dollars en 2013 (source : rapport HCSP).

1.2.1.3 Une dynamique importante de création d'entrepôts de données de santé, majoritairement par les établissements de santé

L'annexe 1 présentant un panorama des entrepôts de données de santé détaille les constats synthétisés dans cette partie.

Selon la Haute Autorité de santé, « les entrepôts de données de santé (EDS) désignent la mise en commun des données d'un ou plusieurs systèmes d'information médicaux, sous un format homogène pour des réutilisations à des fins de pilotage, de recherche ou dans le cadre des soins ».

La CNIL a autorisé 63 entrepôts de données entre 2017 et mi 2023, avec une dynamique croissante³⁸. Une faible proportion des autorisations concerne un entrepôt chaîné avec le SNDS (9 autorisations relevant de 3 acteurs, l'INCA, Clinityx et Agoria). Notons toutefois que la notion même d'entrepôts de santé est d'interprétation variable, et qu'une clarification de la notion aurait du sens.

Les entrepôts de données de santé autorisés émanent en majorité (55 %) des établissements de santé (publics, établissements non lucratifs en particulier de cancérologie et, plus récemment, établissements lucratifs). Près de 30 % des autorisations d'entrepôts de données de santé émanent d'entreprises privées le plus souvent spécialisées dans l'analyse de données. Enfin, le reste des autorisations relève d'acteurs très variés : sociétés savantes, réseaux de soins, structures académiques (par ex INSERM), etc.

Cette dynamique de structuration des entrepôts de données de santé, particulièrement dans le champ hospitalier, est un atout important pour l'enrichissement du patrimoine de données de santé français. Elle a été soutenue par les pouvoirs publics, en particulier avec un appel à projet doté de 75 millions d'euros lancé en 2022.

Au sein du paysage français des bases de données en santé, les entrepôts de données de santé hospitaliers présentent deux avantages.

En premier lieu, **ils intègrent de nombreuses données médicales absentes du SNDS.** Ainsi, d'après le rapport de la Haute Autorité de santé³⁹, tous les EDS intègrent notamment la biologie et les données textuelles, et très souvent les données sur le circuit du médicament qui ne font pas partie du SNDS. Plus rarement, ils intègrent des types de données complexes et volumineuses comme la réanimation, l'imagerie ou la génomique.

En second lieu, **les données des entrepôts de données de santé sont produites en routine**, à partir des systèmes d'information hospitaliers, après une étape de pseudonymisation. Les EDS recouvrent ainsi des données médicales recouvrant un grand nombre de patients (14 millions de patients pour l'entrepôt de AP-HP), largement supérieur aux volumes de patients des cohortes et des registres. Ces derniers disposent toutefois de données médicales d'une qualité et d'une homogénéité généralement supérieure aux données hospitalières et ciblées (sur une population, une pathologie et dans une zone géographique déterminée...).

³⁸ Le nombre total d'entrepôts est toutefois inférieur à 63, certains porteurs de projets ayant sollicité plusieurs autorisations, avec des demandes de modification d'une autorisation initiale.

³⁹ Ce rapport s'appuie sur une enquête menée auprès de 17 CHU et 5 autres établissements hospitaliers.

Les EDS hospitaliers poursuivent trois finalités : **la recherche, le pilotage et le soin (détail en annexe 1)**. Plusieurs interlocuteurs de la mission ont souligné que les entrepôts de données de santé pourraient contribuer encore plus significativement à l'amélioration de la qualité des soins, avec des retours réguliers vers les professionnels de santé sur leurs pratiques, le développement d'outils de recherche transverse, etc. **Ceci appelle à atténuer la distinction aujourd'hui stricte entre utilisation primaire et secondaire des données de santé.**

Les entrepôts de données hospitaliers souffrent toutefois de la **faiblesse de la documentation des systèmes d'information hospitaliers** par les éditeurs de logiciel. Cette faiblesse « à la racine » de la production de donnée se double d'une forte hétérogénéité dans la manière de structurer les données au sein de chaque entrepôt. Ainsi, **les standards de données restent encore peu utilisés ou hétérogènes** selon les centres, et il n'y a pas à ce stade de recours aux interfaces de programmations API pour faciliter les échanges de données (voir point 1.2.2.3).

Deux types d'initiatives ont émergé pour harmoniser les données des entrepôts de données de santé et renforcer la maturité des EDS comme leur interopérabilité.

En premier lieu, le **développement de réseaux d'entrepôts à l'échelle régionale ou interrégionale** (voire nationale, pour la cancérologie) est une tendance marquée, afin de développer des solutions techniques communes, mutualiser les expertises et harmoniser les données. Les dynamiques les plus avancées concernent le Grand Ouest (voir encadré ci-dessous), ou, dans le domaine de la cancérologie, la plateforme ESME, développée par UNICANCER, regroupe les données des centres de lutte contre le cancer.

La structuration interrégionale des entrepôts de données du Grand ouest

Depuis 2016, le groupement de coopération sanitaire (GCS) Hugo a lancé un programme pluriannuel pour structurer et organiser l'exploitation des données en santé sur le territoire du Grand Ouest. Un réseau interrégional a été créé pour structurer les centres de données cliniques créés au sein des 6 établissements du GCS (les CHU de Rennes, Brest, Angers, Nantes, Tours et l'institut de cancérologie de l'ouest), avec un accompagnement financier, et technique et scientifique de projets multicentriques. Ce rapprochement se traduit par l'utilisation de technologies communes et la définition de standards de données pour favoriser l'interopérabilité. Il s'appuie aussi sur la mutualisation de certaines ressources et expertises, et l'adoption de principes communs de gouvernance et de valorisation. En 2020, une plateforme de données commune (Ouest Data hub) a également été créée pour l'intégration de données pour les projets multicentriques et le chaînage des données. Celle-ci utilise l'infrastructure du CHU de Nantes. Le réseau a la capacité de chaîner (de manière décentralisée) ses données avec les données du SNDS.

La mobilisation des données chaînées au SNDS a permis la réalisation d'une étude en vie réelle sur les statines dans le champ de la pharmacovigilance, mesurant la prévalence des interactions médicamenteuses et des effets indésirables liés aux prescriptions au cours du parcours de soin du patient en ville et à l'hôpital. Cette étude a donné lieu à une publication⁴⁰.

⁴⁰ Publication : Bannay, Aurélie, Mathilde Bories, Pascal Le Corre, Christine Riou, Pierre Lemordant, Pascal Van Hille, Emmanuel Chazard, Xavier Dode, Marc Cuggia, et Guillaume Bouzillé. « Leveraging National Claims and

En second lieu, **des travaux ont été engagés à l'échelle nationale, sous l'égide du comité stratégique des données de santé, pour harmoniser les approches des entrepôts de données de santé.** Ainsi, des groupes de travail ont été initiés concernant la gouvernance des entrepôts, l'établissement d'un socle de données, le financement des entrepôts de données de santé ainsi que les principes de tarification (réflexion sur une grille commune). Ces travaux partenariaux constituent une avancée, comme en attestent les premiers résultats en matière de financement présentés *infra*, même s'ils ne vont pas encore jusqu'à l'adoption de normes communes, notamment en termes d'interopérabilité.

1.2.1.4 Un paysage de bases de données foisonnant mais peu lisible, avec certains champs insuffisamment couverts comme la médecine de ville

Au-delà des bases de données précédemment présentées (SNDS, registres, cohortes et entrepôts de données de santé hospitaliers), il existe de nombreuses autres bases de données permettant une utilisation secondaire de données de santé. Sans viser l'exhaustivité, on peut évoquer les bases des essais cliniques, les bases de données issues des systèmes d'information des agences sanitaires (par exemple, la base de données CRISTAL de l'agence de biomédecine issue du système de suivi des prélèvements et des greffes d'organes et de tissus), les enquêtes (comme l'enquête annuelle FLAHS pour le suivi de la prévalence de l'hypertension en France), et panels statistiques ou encore les recueils spécifiques sur les études observationnelles (comme les études cas-témoins pour étudier les facteurs de risque d'une maladie rare). De nouvelles sources de données devraient émerger, par exemple les données issues des objets connectés des patients.

Ce paysage se caractérise par de nombreuses initiatives, caractéristique d'un secteur en pleine évolution technologique. Là encore, sans viser l'exhaustivité, on peut citer :

- Bien sûr, la constitution par le Health data hub d'une plateforme technologique, d'un catalogue de données et de nombreuses initiatives (détail des réalisations en partie 1.3.1) ;
- La constitution d'infrastructures plus spécialisées, comme France Cohortes (encadré ci-après), ou le projet de collecteur analyseur de données (CAD) pour les données génomiques ;
- L'émergence de nombreuses collaborations entre acteurs. La plateforme de cancérologie de l'INCA, co-financée par la filière d'intelligence artificielle et cancers (FIAC), constitue un exemple d'une telle collaboration, avec un investissement de 9 millions d'euros par les industriels (financement abondé de 8 millions d'euros par la BPI). Le consortium Agoria⁴¹ constitue un autre exemple de rapprochement entre acteurs (privés), afin de constituer une infrastructure de données de santé destinée à héberger des données appariées avec le SNDS, avec l'ambition d'héberger un système fils du SNDS.

Ces différentes initiatives pâtiennent toutefois d'un manque d'articulation (cf partie 1.3.3).

Hospital Big Data: Cohort Study on a Statin-Drug Interaction Use Case ». JMIR Medical Informatics 9, no 12 (13 décembre 2021): e29286. <https://doi.org/10.2196/29286>.

⁴¹ Le consortium, dont le chef de file est Docaposte, compte aussi parmi ses membres Astrazeneca, Impact Healthcare et Takeda. Il n'a pas de personnalité morale à ce jour.

France Cohortes

France Cohortes est une infrastructure mutualisée pour les cohortes en cours de développement par l'INSERM, bénéficiant de crédits du programme Investissement d'avenir. Elle comprend plusieurs volets :

- une plateforme informatique commune, conforme au référentiel de sécurité du SNDS et au RGPD, hébergé par l'INSERM dans son *data center* situé au Centre Informatique national de l'enseignement supérieur de Montpellier (CINES). La plateforme vise à assurer la gestion informatique des cohortes (matériels, logiciels, réseaux, installations, maintenance), de manière interopérable avec d'autres systèmes d'information, avec bulles sécurisées permettant aux chercheurs de traiter les données de manière sécurisées. Son entrée en services est prévue en janvier 2024 ;
- une animation scientifique des cohortes et le développement de leur visibilité, à travers la mise en place d'un catalogue de données et la simplification de la gouvernance d'accès aux données (mise en place d'un outil de dépôt de demande, et d'outils de mise à disposition des données) à partir de janvier 2024 ;
- le développement de l'interopérabilité, avec des travaux sur les modèles de données (standardisation de la donnée à la source).

Un accompagnement juridique et réglementaire.

Neuf cohortes ont d'ores et déjà rejoint la démarche, avec une migration de données en cours, un conventionnement prévu au 31/12/23, et une mise au catalogue au premier semestre 2024. Onze bases de données de maladies rares (financées par l'appel à manifestation maladies-rares) font également partie de Frances cohortes.

La contrepartie de ce foisonnement est aussi un manque de lisibilité. En dépit de l'existence de premiers portails de recensement des bases de données, comme le portail Epidémiologie France, les acteurs économiques en particulier soulignent que le paysage des bases de données est complexe, éparpillé et hétérogène, et que le contenu précis des bases de données et de leur qualité est souvent méconnu, ce qui nuit à la réutilisation des données.

Par ailleurs, il existe plusieurs manques importants dans les bases de données en santé. Sans être exhaustif, quelques illustrations peuvent être apportées.

En premier lieu, la faiblesse de la structuration des données de ville, à l'exception des panels de médecins de ville développés par plusieurs entreprises privées, est déplorée par de très nombreux interlocuteurs. Le projet d'entrepôt de données de médecine de ville porté par le conseil national des généralistes enseignants (CNGE), accompagné par le Health data hub et soutenu par le programme d'investissement France 2030 constitue ainsi un enjeu important.

Le projet de création d'un entrepôt de données de médecine de ville (P4DP)

Ce projet est porté par un consortium composé du Conseil national des généralistes enseignants (CNGE), de la plateforme des données de santé (Health data hub), du CHU de Rouen, de l'université de Nice, de l'université de Normandie et de l'entreprise OMICS. Il s'appuie sur de premières expériences d'entrepôts de médecine de ville en région PACA puis en région Normandie.

Il bénéficie d'un financement de 9 millions d'euros dans le cadre de l'AMI Santé numérique, et d'un accompagnement par le Health data hub. La plateforme, qui devrait être opérationnelle début 2024, vise à regrouper les données des dossiers médicaux informatisés des médecins généralistes à partir des données de leurs logiciels de gestion. L'interopérabilité avec les logiciels sera intégrée au cahier des charges soumis aux éditeurs de logiciel dans le cadre du programme Ségur du numérique en santé.

Plusieurs services proposés doivent favoriser l'adhésion des médecins, avec un objectif de 2000 recrutements en 3 ans :

- un retour d'informations sur leurs pratiques ;
- une veille épidémiologique ;
- un outil au service de la recherche en médecine de ville (notamment pour l'identification des patients pour intégrer des protocoles).

Les données de l'entrepôt seront versées au catalogue du Health data hub (avec un appariement permanent prévu avec le SNDS), à l'exception à ce stade des données textuelles brutes.

En second lieu, certains champs sectoriels ou géographiques sont peu ou pas couverts par les bases de données françaises. Par exemple, en matière de registres, plusieurs des thématiques prioritaires de la politique de santé publique ne sont pas « couvertes » par des registres comme le diabète, le surpoids et l'obésité, la santé mentale ou la santé sexuelle. Le manque de bases de données relatives aux maladies chroniques a été déploré par certains interlocuteurs de la mission.

De même, certaines zones géographiques sont moins couvertes, comme les départements et territoires d'Outre-mer.

1.2.2 Des bases de données encore sous-exploitées et peu appariées, dont l'interopérabilité n'est pas suffisamment pensée *ab initio*

1.2.2.1 Un positionnement de la France relativement en retrait par rapport à d'autres pays étrangers

Le nombre de publications liées aux bases de données françaises reste largement inférieur à certaines bases étrangères très accessibles. Au-delà de la cohorte britannique UK Biobank déjà mentionnée, on peut citer la base américaine MIMIC concernant des données de réanimation sur 40 000 patients, qui a fait l'objet de 1461 publications en 2023 (à date d'octobre) ou la base VAERS/FAERS de pharmacovigilance de la FDA (1515 publications en 2023 à date d'octobre). Même si les contextes réglementaires sont très différents et que le volume de publications n'est

qu'un indicateur parmi d'autres, ces données illustrent que le potentiel des bases de données françaises est encore sous-exploité.

D'après une note publiée par le Health data hub sur la valorisation de l'utilisation secondaire des données de santé⁴² (2022), « les travaux de l'OCDE⁴³ indiquent que **la France se situe plutôt à un haut niveau sur les questions de gouvernance, et à un niveau plus moyen sur les questions de maturité et d'usage secondaire des données.** »⁴⁴ D'après l'OCDE, ce positionnement en retrait s'explique également par l'absence de bases de données sur certaines aires thérapeutiques (par exemple le diabète ou les maladies cardio-vasculaires), par le manque d'automatisation dans la collecte des données, notamment cliniques, ou par des facteurs plus structurels (temps pour collecter les données, difficultés d'identification unique entre les bases, faible utilisation pour évaluer la performance).

Il y a consensus pour constater que **l'attractivité internationale de la France repose sur des données médico-administratives particulièrement riches** (la base principale du système national des données de santé et en particulier le SNIIRAM), dont l'accès pâtit toutefois d'une durée d'accès très longue qui entrave nettement sa compétitivité à l'échelle internationale (cf. partie 1.2.3). Par ailleurs, la structuration de données médicales dans les entrepôts de données (notamment hospitaliers) est reconnue, mais leur qualité reste encore perfectible et méconnue. De manière générale, les partenariats peuvent être longs à conclure, avec une sensibilité particulière sur les questions de valorisation (cf. point 1.3.4).

Dans ce contexte, un certain nombre d'entreprises et de chercheurs privilégient l'utilisation de données à l'étranger, auprès de pays très compétitifs dans ce domaine, qui disposent toutefois de réglementations très différentes.

⁴² "Quelle valorisation pour l'utilisation secondaire des données de santé ? », Veltys, 2022

⁴³ OECD Health Working Paper No. 127 SURVEY RESULTS: NATIONAL HEALTH DATA INFRASTRUCTURE AND GOVERNANCE Jillian Oderkirk, 2021

⁴⁴ La gouvernance se réfère à l'ensemble des organisations et procédures mises en place dans le pays afin d'encadrer la collecte des données et leur utilisation, quand la disponibilité, la maturité et l'usage dans l'utilisation des données sont calculés sur 13 bases représentatives de l'ensemble des données de santé. Ce dernier item valorise la disponibilité de données cliniques plutôt que de données médico-administratives, ce qui contribue à un positionnement en retrait de la France.

Exemples étrangers en matière d'accès à la donnée

Les **Etats-Unis** se classent parmi les pays avec l'une des meilleures gouvernances de l'OCDE en termes de données de santé, avec un effort important dans la mise à disposition des données pour la recherche. Une autre de leurs caractéristiques tient à l'autorisation pour les organismes de santé d'exploiter les données qu'ils collectent et de donner ou vendre les données (pseudonymisées) à des partenaires pour la recherche, sans consentement ni information des patients. Ceci contribue à ce que les entreprises américaines soient aujourd'hui les mieux valorisées au monde (voir encadré précédent), même si les Etats-Unis disposent de peu de bases de données nationales exhaustives (à l'exception des National Institutes of Health) en raison d'un système pluriel et fondé sur les assurances privées.

Le **Royaume-Uni** a mis en place une gouvernance nationale des données, avec le Health data research UK, qui fédère 86 organisations pour gérer la gouvernance et la qualité des données au niveau national, avec une politique très proactive de partage des données. Il dispose également de plusieurs banques de données reconnues mondialement, comme des cohortes nationales (UK Biobank, qui regroupe 500 000 participants entre 40 et 69 ans) ou des données de soins primaires comme le Clinical Practice Research Datalink (environ 16 millions de patients), ou encore les données du National Health System.

Israël a un positionnement très compétitif en termes de réutilisation des données de santé, en raison de la couverture de l'ensemble de la population par quatre acteurs assurantiels, avec une numérisation ancienne des dossiers patients (données cliniques avec une profondeur de 25 ans). Il dispose d'un écosystème de start-up très dynamique, avec d'importants investissements en R&D en santé numérique, notamment en télémédecine, aide à la décision et diagnostic.

Source : Extraits de la note "Quelle valorisation pour l'utilisation secondaire des données de santé ? », Veltys, 2022

1.2.2.2 Des appariements encore peu développés

Les bases de données françaises restent peu appariées entre elles, alors que les analyses croisant plusieurs bases apportent une plus-value importante. Les travaux réalisés par la DREES sur l'EDP Santé⁴⁵, croisement des données du SNDS et des données socio-fiscales, illustrent le fort potentiel d'exploitation de données appariées. On a aussi vu précédemment que les données des cohortes et registres gagneraient à être enrichies de données d'autres bases de données (SNDS, entrepôts de données mais aussi données socio-fiscales), pour conforter la robustesse des données ainsi que pour automatiser l'intégration de certaines informations.

Mais de tels appariements restent rares et particulièrement complexes à mener. A titre d'illustration, parmi les études portant sur le SNDS publiées en 2022, moins de 10% étaient des

⁴⁵ L'EDP-Santé croise les données de l'échantillon démographique permanent (EDP), produit par l'Insee, et celles du SNDS. L'exploitation de cette base de données a donné lieu à deux études illustrant l'impact du revenu sur la santé : « Les maladies chroniques touchent plus souvent les personnes modestes et réduisent davantage leur espérance de vie » ; Études et Résultats n° 1243, octobre 2022 ; et « En France, les AVC sont plus fréquents, plus graves et moins souvent pris en charge en unité spécialisée pour les personnes les plus modestes » Études et Résultats n° 1219 février 2022

études chaînées (essentiellement à des registres ou cohortes). De même, l'établissement français du sang a témoigné avoir dû renoncer à un projet d'appariement pour démontrer l'innocuité de la plasmaphérèse, qui présente pourtant un impact fort en termes de santé publique et d'image, en raison de blocages techniques et conventionnels.

Le faible développement des appariements s'explique par plusieurs facteurs :

- le cadre réglementaire rend ces appariements complexes, en limitant fortement la collecte du numéro de sécurité sociale (numéro d'inscription au répertoire, par la suite désigné NIR) dans le cadre d'une recherche (soumis à une autorisation préalable de la CNIL), en encadrant les appariements avec les données du SNDS (autorisation de la CNIL nécessaire, y compris pour les acteurs avec un accès permanent), et en requérant l'information individuelle des patients pour chaque appariement⁴⁶. La méconnaissance du cadre réglementaire (qui a beaucoup évolué ces dernières années) contribue aussi sans doute à limiter les appariements.
- l'ensemble des acteurs (acteurs institutionnels, acteurs de la recherche mais aussi acteurs privés) ont témoigné de la longueur et de la difficulté des procédures d'appariements, que ce soit des appariements avec les données de la base principale du SNDS (cf. partie 1.2.2.2. s'agissant des difficultés spécifiques d'accès à la base principale du SNDS), mais aussi les appariements des données entre EDS hospitaliers, ou encore les appariements avec des bases de données d'autres champs (exemple des données socio-professionnelles de la CNAV ou des données fiscales).
- les appariements sont rendus complexes par la faible interopérabilité des bases de données. Comme mentionné précédemment, les entrepôts de données de santé utilisent des normes différentes, et les registres utilisent peu de codifications internationales. Si les bases de données ne sont pas conçues avec cette logique d'interopérabilité, celle-ci est rendue beaucoup plus complexe *a posteriori*.

1.2.2.3 Les API, une opportunité insuffisamment exploitée pour le partage des données

Le recours aux interfaces de programmation applicatives, ou « API » (« *Application Programming Interface* ») est un vecteur important pour faciliter les échanges de données entre systèmes d'information. Contrairement à ce qui est mis en place dans d'autres secteurs technologiques, les travaux de standardisation et d'interopérabilité des Données de la santé en France semblent ainsi encore très insuffisants pour pouvoir s'appuyer sur des API pour industrialiser et automatiser ces échanges de manière sécurisée.

Le HDH confirme ne pas exposer d'API sur sa plateforme technologique, n'envisageant à ce stade d'y recourir que pour exposer les métadonnées de son catalogue. L'AP-HP indique néanmoins s'être engagée, pour faciliter l'exploitation secondaire des données de son EDS, dans l'installation d'une modélisation permettant à la fois l'intégration des données en amont et leur mise à

⁴⁶ Dans les faits, la CNIL accepte généralement la dérogation à cette information individuelle justifiée par des questions d'impossibilité ou d'insuffisance de moyens.

disposition en aval à différents types de clients par voie d'API pour différents types d'usage (dont la recherche – incluant les méthodologies modernes de type IA –, le pilotage, l'appui aux soins).

Cette absence de recours aux interfaces logicielles entre systèmes d'information et bases de données du secteur de la santé en France paraît regrettable au regard des pratiques depuis 20 ans dans les autres secteurs, y compris dans le secteur de la santé, dans d'autres pays. Ainsi, aux Etats-Unis, la loi "21st Century Cures Act", signée en 2016, exige que les systèmes d'information en santé disposent d'une interface de programmation d'application (API) donnant accès à tous les éléments de données du dossier de santé électronique d'un patient, "sans effort particulier".

En matière de santé, des travaux internationaux de standardisation des formats de données, portés en France par l'Agence du numérique en santé, sont en cours depuis plusieurs années (IHE, HL7, FHIR) y compris pour l'échange massif de données de santé par API (the SMART/HL7 FHIR Bulk Data Access Application Programming Interface)⁴⁷.

Le recours à terme à des API n'est pas simplement une commodité technique pour faciliter les échanges. Il exige une mobilisation de toute la chaîne du numérique en santé jusqu'aux éditeurs de logiciels présents dans les systèmes d'information des acteurs en santé pour rentrer pleinement dans une dynamique réelle de partage des données de santé, documenter les bases de données, s'entendre sur des standards de données pour pouvoir à terme développer des API.

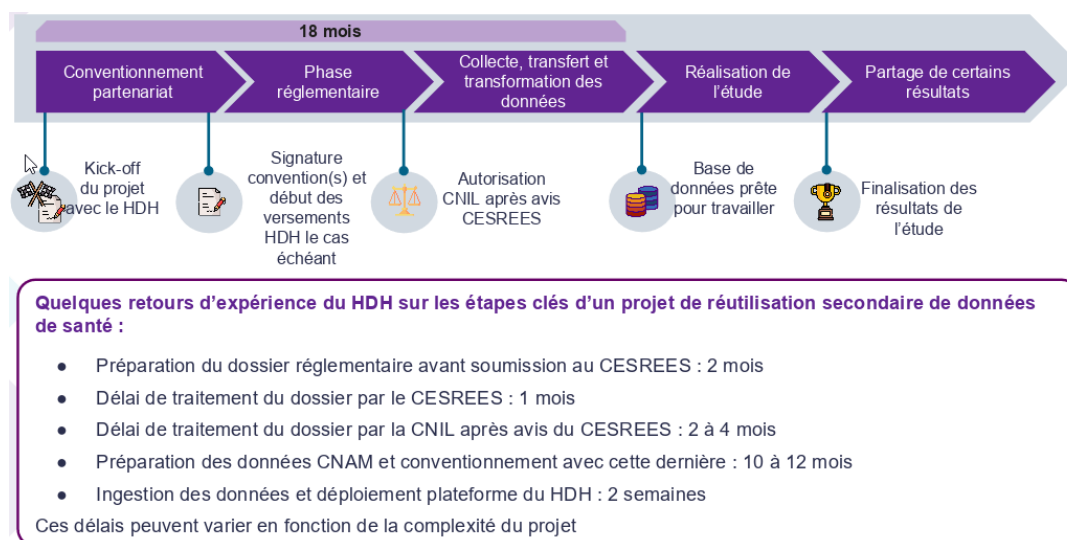
1.2.3 Un accès aux données de santé qui s'est allongé malgré la création de la plateforme des données de santé (Health Data Hub)

1.2.3.1 Des délais d'accès aux données de santé importants, en particulier s'agissant de la base principale du SNDS

En France, le parcours d'accès standard aux données de santé de la base principale du SNDS se caractérise par sa **complexité** et sa **lourdeur** tant au niveau de la phase réglementaire qu'au niveau du ciblage et de l'extraction des données requêtées. À l'heure actuelle, le délai entre la soumission du projet de recherche mobilisant des données de santé et la mise à disposition effective de ces données est évalué, en moyenne, à **18 mois**, comme l'illustre ce schéma.

⁴⁷ <https://doi.org/10.1038/s41746-020-00358-4>.

Schéma 1 : Les grandes étapes d'un projet de recherche à partir de données de santé en France



Source : Health Data Hub

Les délais d'accès sont également importants lorsque sont requêtées d'autres bases de données que la base principale du SNDS. À titre d'exemple, la cohorte Constances indique que s'écoulent en moyenne 53 semaines (au minimum 36 semaines) entre le dépôt de la demande d'un chercheur n'appartenant pas à l'INSERM et la mise à disposition des données demandées de la cohorte, ce délai pouvant varier selon l'établissement auquel appartient le responsable de traitement. Au-delà des démarches réglementaires, l'étape de contractualisation pour la mise à disposition des données entre l'INSERM et l'établissement concerné est particulièrement chronophage (entre 21 et 43 semaines).

De même, un projet multicentrique porté par une chercheuse de l'AP-HP, visant à collecter des données rétrospectives et prospectives à partir de quatre grands CHU et quatre hôpitaux de l'AP-HP n'a pu obtenir l'aval de la CNIL qu'au bout d'un an pour que le projet puisse être hébergé par le HDH. Si la complexité du projet, qui implique de travailler sur des données d'imagerie dans un format natif, peut en partie expliquer cette lenteur réglementaire, celle-ci reste particulièrement pénalisante dans un domaine de recherche compétitif à l'échelle nationale et internationale.

Enfin, certains porteurs de projet entendus par la mission ont reconnu que leur propre institution de rattachement pouvait faire preuve de résistance dans la coopération avec le HDH, rendant d'autant plus délicate l'étape de conventionnement.

1.2.3.2 Un environnement réglementaire très complexe malgré l'existence de procédures simplifiées

Compte-tenu de la mission confiée en juillet 2023 au comité consultatif national d'éthique pour les sciences de la vie et de la santé, portant sur l'organisation de l'évaluation éthique pour les recherches impliquant la personne humaine, la mission n'a pas traité des procédures sur ce champ. Cette partie concerne ainsi spécifiquement les demandes d'autorisation ou d'études n'impliquant pas la personne humaine.

a) Des conditions légales d'accès aux données de santé relativement exigeantes par rapport aux autres pays européens

L'article 89, paragraphe 1 du RGPD soumet les traitements à des fins de recherche de données à caractère personnel sensibles, dont les données de santé, à des « **garanties appropriées pour les droits et libertés de la personne concernée** » qui doivent garantir « *la mise en place de mesures techniques et organisationnelles, en particulier pour assurer le respect du principe de minimisation des données* ». Au titre de ces « garanties appropriées », la législation française exige des projets de recherche recourant à des traitements de données de santé à caractère personnel qu'ils répondent à une **finalité d'intérêt public** et les soumet, dans le cadre de l'accès standard (hors procédures simplifiées), à une **autorisation préalable délivrée par la CNIL**, selon des modalités décrites en annexe 2 du présent rapport. Le législateur français a maintenu un régime d'autorisation par une autorité pour les recherches nécessitant un accès à des données de santé, quand plusieurs autres pays européens font reposer la responsabilité de l'octroi de cet accès sur le responsable de la base de données requêtée, moyennant, le cas échéant, l'avis préalable d'un comité scientifique et/ou d'éthique (cf. annexe 3 sur les comparaisons internationales).

b) Des délais règlementaires d'accès globalement respectés pour les trois quarts des demandes d'autorisation

Le Health data hub assure le secrétariat unique du dépôt de ces projets lorsqu'ils n'impliquent pas la personne humaine et les transmet pour avis au comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé (CESREES) qui se prononce, dans un délai d'un mois renouvelable une fois, sur la méthodologie retenue et sur la pertinence éthique et scientifique du projet, et, le cas échéant, de sa propre initiative ou sur saisine du président de la CNIL ou du ministre chargé de la santé, sur son caractère d'intérêt public.

Depuis 2020, soit dans sa configuration rénovée issue du décret n° 2019-536 du 29 mai 2019⁴⁸, le CESREES examine les projets de traitement de données de santé à caractère personnel à des fins de recherche n'impliquant pas la personne humaine à raison d'une réunion par mois. Au 28 août 2023, il avait rendu 1 233 avis concernant 906 projets, dont 522 projets impliquant un accès à des données du SNDS, pour lesquels 717 avis ont été rendus (un peu plus de 58 % du total des avis rendus depuis 2020).

Ce sont donc 36 % des projets soumis au CESREES qui font l'objet d'un réexamen. Près de 60 % des projets obtiennent un avis favorable sans avoir à repasser devant le CESREES et, au total, 84 % des projets sont acceptés par le CESREES, le cas échéant après modification en vue d'un réexamen. Les motifs de refus opposés par le CESREES tiennent le plus souvent à un problème de méthodologie, certains projets se fixant notamment pour objectif d'obtenir du SNDS des informations qui, en réalité, n'existent pas.

Certains acteurs de la recherche regrettent l'absence de transparence et de visibilité sur la doctrine du CESREES dans la caractérisation d'une finalité d'intérêt public ; d'autres font état d'un manque de constance dans l'évaluation scientifique de projets pourtant similaires par les experts du CESREES. Pour sa part, le CESREES indique analyser l'intérêt public par le prisme de

⁴⁸ Pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

l'amélioration de la prise en charge des citoyens, laquelle n'est *a priori* pas incompatible avec l'intérêt commercial de l'étude⁴⁹.

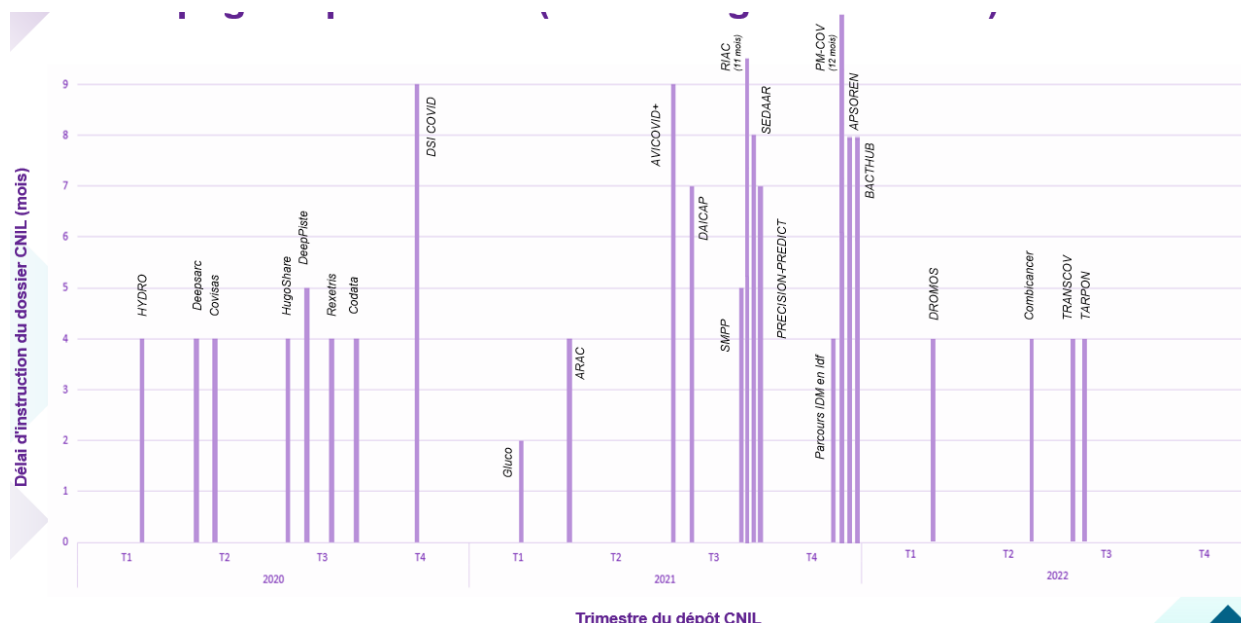
Une fois obtenu l'avis favorable du CESREES, la demande est transmise par le HDH à la CNIL qui doit examiner la conformité du projet de recherche aux règles du RGPD ainsi qu'aux dispositions de la section 3 du chapitre III du titre II de la loi « Informatique et libertés » du 6 janvier 1978, dite loi « Informatique et libertés », et du titre VI du livre IV de la première partie du code de la santé publique. La CNIL dispose alors d'un délai de deux mois, renouvelable une fois, pour délivrer l'autorisation, laquelle peut être tacite en l'absence de décision expresse à l'expiration de ce délai.

En pratique, **les porteurs de projet se satisfont rarement d'une décision d'acceptation tacite** et sollicitent donc, par souci de sécurité juridique, une confirmation écrite de la CNIL d'autorisation de mise en œuvre du traitement à l'issue du délai légal. Si la CNIL reconnaît qu'il arrive qu'une décision expresse d'autorisation intervienne dans un délai de plus de quatre mois, cette situation ne concernerait que des autorisations – aucune décision de refus n'ayant été rendue postérieurement au délai légal – relatives le plus souvent à des dossiers qui restent en cours de finalisation à l'expiration du délai, notamment dans l'hypothèse où la CNIL aurait demandé des compléments ou précisions sur des points mineurs ou non bloquants du projet, ou à des dossiers qui, soit en raison de leur qualité insuffisante, soit de leur complexité particulière, ont nécessité des demandes d'informations complémentaires.

Dans ces conditions, si la CNIL respecte généralement le délai légal maximum de quatre mois pour autoriser ou refuser d'autoriser un traitement de données de santé à des fins de recherche, les délais d'autorisation de certains projets ont pu largement dépasser les 4 mois réglementaires, pour atteindre, dans certains cas, près d'un an :

⁴⁹ Il exige des projets que leurs finalités soient exposées clairement et mettent en évidence le bénéfice attendu sur les individus ou leur santé, ou en termes d'amélioration de la qualité du système de santé. À cet égard, l'engagement des porteurs de projets, notamment industriels, à publier les résultats de leur recherche peut constituer une garantie de l'intérêt public, le CESREES se gardant la possibilité de ne donner qu'un avis réservé tant qu'une réponse positive claire n'a pas été apportée par le porteur de projet sur ce point.

Graphique 2 : Délais d'instruction par la CNIL depuis 2020 pour les projets accompagnés par le HDH



Source : Health Data Hub.

On constate, sur la période récente – à l’exception notable de 2020 –, que la proportion de dossiers ayant fait l’objet d’une autorisation expresse de la CNIL par décision déléguée, incluant les demandes de traitement de données pour des non-RIPH et des RIPH, délivrée au-delà du délai réglementaire de quatre mois se maintient autour d’un **quart du total des dossiers autorisés**.

Tableau 2 : Proportion de demandes de traitement de données de santé délivrée par la CNIL au-delà du délai de 4 mois entre 2020 et 2023

	2020	2021	2022	2023 (au 21 août)
Nombre d'autorisations de traitement RIPH et non-RIPH accordées par décisions déléguées ⁵⁰	370	352	274	180
Nombre de refus d'autorisation par décisions déléguées	30	12	12	-
Proportion de dossiers (RIPH et non-RIPH) autorisés par décisions déléguées au-delà du délai de 4 mois	11 %	22 %	28 %	24 %

Source : CNIL.

⁵⁰ La plupart des autorisations de traitement accordées par la CNIL interviennent par décisions déléguées, le reste faisant l’objet d’une délibération en séance plénière.

c) Des démarches réglementaires jugées complexes et parfois dissuasives

De nombreux acteurs auditionnés par la mission, font valoir que, si la richesse des bases de données de santé en France, et en particulier de la base principale du SNDS, est de plus en plus reconnue au niveau européen et international, **la complexité des démarches réglementaires et les délais d'accès continuent d'exercer un effet dissuasif qui peut conduire les équipes de recherche ou les acteurs économiques à renoncer à mobiliser les données de santé françaises.**

Sont souvent cités à titre de comparaisons internationales, les États-Unis, où un accès direct à des données de santé peut être obtenu en moyenne en une semaine dans le cadre de licences d'accès, et le Royaume-Uni, où l'accès aux principales bases de données nationales n'est pas subordonné à une autorisation délivrée par une agence de protection des données, mais est délivré à la discrétion de l'institution détentrice de la base, selon un processus d'examen interne qui lui est propre et faisant intervenir une analyse scientifique et éthique des demandes adaptée au degré de complexité de celles-ci (cf. annexe 3 sur les comparaisons internationales).

Enfin, la phase réglementaire de l'accès aux données est marquée par une possible **redondance de l'examen d'un projet de recherche par plusieurs comités**. À titre d'illustration, un projet mobilisant des données d'une cohorte impliquant un chaînage *ad hoc* à la base principale du SNDS nécessitera non seulement l'accord du comité scientifique et éthique de la cohorte requêtée et l'accord du CESREES, mais également, le cas échéant, l'accord du comité scientifique et éthique de l'étude et de celui de l'institution à laquelle appartient le porteur du projet. À cela peut s'ajouter l'accord d'un comité éthique bénéficiant de la qualification de l'« *International Review Board* » (IRB) afin de permettre la publication des résultats de l'étude dans une revue internationale. En définitive, le projet de recherche aura potentiellement fait l'objet de multiples passages devant des comités différents, occasionnant une charge disproportionnée pour l'équipe de recherche concernée et un allongement significatif des délais.

D'une façon générale, il convient de souligner un sentiment prégnant chez les chercheurs d'être insuffisamment compris des équipes juridiques et des délégués à la protection des données (DPO) de leur institution de rattachement, qui sont perçus comme faisant preuve d'excès de précaution dans la mise en œuvre des démarches réglementaires. Les organismes de recherche soulignent, à cet égard, la difficulté à acculturer leurs équipes juridiques à l'environnement réglementaire dans la recherche à partir des données de santé et à attirer et recruter des personnes expertes en droit du numérique en santé.

d) Des procédures simplifiées qui fluidifient insuffisamment le système

En alternative à l'autorisation préalable de la CNIL, peuvent être mobilisées des procédures d'accès simplifiées :

- la **déclaration de conformité à une des méthodologies de référence** élaborées par la CNIL en concertation avec le HDH et des organismes publics et privés intéressés (détaillées en annexe 2) ;
- la **décision unique** accordée à un même responsable de traitement (il s'agit, en pratique, essentiellement de bureaux d'études) pour la réalisation de plusieurs traitements à finalités, catégories de destinataires et catégories de données identiques ;

- le dispositif des accès permanents qui permet à des organismes, listés à l'article R. 1461-12 du code de la santé publique (CSP), de traiter des données de santé à caractère personnel issues de la base principale du SNDS, sans possibilité d'appariement.

Tableau 3 : Nombre de déclarations de conformité effectuées par année

	MR-001	MR-002	MR-003	MR-004	MR-005	MR-006
2019	982	31	731	715	153	49
2020	814	16	741	1 021	91	31
2021	930	18	390	652	82	34
2022	957	26	390	652	65	23
2023 (au 01/09)	535	11	338	657	44	14

Source : CNIL.

Sur 1 026 projets de traitements de données de santé à des fins de recherche, d'étude ou d'évaluation « autorisés » n'impliquant pas la personne humaine en 2022, 286 ont fait l'objet d'une autorisation de la CNIL (dont 5 autorisations tacites) et 740 ont mobilisé une méthodologie de référence, soit un **rapport de 28 %/72 %**. En 2021, le rapport entre la part de traitements « recherche » autorisés par la CNIL et la part de traitements mobilisant une méthodologie de référence était de 32 %/68 %.

Il y a donc une **forte mobilisation des dispositifs simplifiés, qui pourraient toutefois être encore mieux appropriés**.

De l'avis général, l'accès permanent est un élément de facilitation considérable pour les acteurs qui en disposent. Toutefois, certains titulaires d'un accès permanent, y compris les agences sanitaires, maîtrisent encore mal les potentialités de cet outil d'exploitation des données du SNDS, dès lors que certaines demandes d'autorisation d'étude de « routine » soumises à la CNIL et ne nécessitant pourtant aucun appariement ni extraction de données auraient pu être satisfaites par la mobilisation de cet accès permanent. À cet égard, un accompagnement renforcé du HDH, lui-même titulaire d'un accès permanent au SNDS, dans la pédagogie de l'utilisation de ce dispositif serait bienvenu.

De même, la CNIL fait observer que le dispositif de la décision unique, qui permet de ne pas avoir à resolliciter une autorisation en cas d'identité de porteur et de finalités, reste très peu mobilisé par les acteurs de la recherche, ses principaux utilisateurs restant les bureaux d'études qui entendent se constituer un patrimoine de données exploitables dans différents domaines de recherche thérapeutique et le valoriser auprès d'acteurs publics et privés de la recherche.

Par ailleurs, si, aux termes de la loi⁵¹, les méthodologies de référence doivent être élaborées par la CNIL en concertation avec le HDH et des représentants des acteurs concernés, **le niveau de co-construction effective de ces méthodologies est jugé perfectible par un certain nombre d'acteurs**, qui regrettent une production de méthodologies de référence insuffisamment réactive,

⁵¹ En application du II de l'article 66 de la LIL et du 5° de l'article L. 1462-1 du code de la santé publique.

ne répondant pas pleinement aux besoins du secteur, ainsi qu'une tendance à assortir ces méthodologies d'exigences qui dépasseraient le cadre réglementaire applicable.

À cet égard, la publication en octobre 2023 des deux nouvelles méthodologies de référence MR-007 et MR-008, destinées à faciliter l'accès à la base principale du SNDS pour des organismes publics et privés, a reçu un accueil mitigé (cf. annexe 2).

e) Une application stricte des principes d'anonymisation et minimisation des données

Enfin la recherche publique et privée en France doit aujourd'hui travailler sur des données appauvries, dans le respect des deux principes d'anonymisation et de minimisation des données personnelles énoncés par le règlement général sur la protection des données (RGPD). Les équipes projets doivent renoncer à utiliser toute la profondeur des données disponibles du fait de l'application stricte de ces deux principes, y compris dans les phases de recherche les plus amont.

Le RGPD prévoit pourtant, à son article 89, la possibilité de déroger au régime général de protection des données à des fins de recherche scientifique, lorsque ce dernier risquerait de "rendre impossible ou d'entraver sérieusement la réalisation des finalités spécifiques et où de telles dérogations sont nécessaires pour atteindre ces finalités". Les termes des dispositions du droit national transposant cet article (article 78 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés et article 116 du décret n° 2019-536 du 29 mai 2019 pris pour l'application de cette loi) restent néanmoins à un niveau assez général et ne semblent pas créer d'assouplissement particulier.

1.2.3.3 Des délais de mise à disposition des données décourageants : le goulot d'étranglement à la caisse nationale d'assurance maladie pour la base principale du SNDS

a) Un délai de mise à disposition en moyenne de 10 à 12 mois

Au-delà des délais réglementaires, **les principales difficultés d'accès aux données de santé à des fins de recherche se nouent à l'étape de la mise à disposition par la caisse nationale d'assurance maladie (CNAM) des données requises.** La préparation des données demandées et leur extraction par la CNAM ainsi que le conventionnement du porteur de projet avec cette dernière nécessitent en moyenne **10 à 12 mois**. Selon des données du HDH, **le délai médian d'accès aux données de la base principale du SNDS aurait augmenté de plus de 56 % entre 2020 et 2022, pour s'établir à 169 jours.** Ce constat doit certes être nuancé par le fait qu'il y a eu en parallèle beaucoup plus d'accès effectifs aux données de la bases principale (cf. *infra*), mais même avec cette nuance la situation reste préoccupante.

La première difficulté réside dans le **manque de moyens à disposition de la CNAM pour absorber le flux des demandes d'accès aux données du SNDS.** Actuellement, la CNAM dispose d'une équipe dédiée à la mobilisation du SNDS à des fins de recherche, comportant 12 personnes⁵². Si les

⁵² Dont 6 statisticiens et *data managers* chargés de réaliser les opérations d'extraction et les ciblage ou appariements correspondant à ces extractions. Au sein de cette équipe, d'autres personnes sont également mobilisées pour le conventionnement avec les porteurs de projet et la gestion des espaces projets.

membres de cette équipe dédiée ont accompagné 37 projets en 2018, ce nombre a été porté à 100 en 2020 et 2021, pour atteindre 165 en 2022, à la suite d'un plan d'actions engagé par l'assurance maladie, une quarantaine de projets restant « en stock ».

Cette équipe serait selon la CNAM désormais au maximum de ses capacités, y compris pour des raisons techniques de systèmes d'information saturés. En effet, **un seul serveur existe pour accueillir toutes les demandes d'accès transitant par le dispositif des accès permanents** – ce qui correspond potentiellement à un public de 2 500 utilisateurs –, ce même serveur étant celui que la CNAM utilise pour ses propres accès permanents pour l'exécution de ses missions. Ce serveur a ainsi été saturé au printemps 2023.

De manière compréhensible, la CNAM n'entend pas investir dans des capacités humaines et techniques supplémentaires pour assurer une mission de mise à disposition des données de la base principale du SNDS à la recherche, **mission que le législateur a confiée au HDH**.

En application de l'article R. 1461-3 du CSP, le HDH et la CNAM sont co-responsables de traitement du SNDS et sont toutes les deux responsables du stockage et de la mise à disposition des données de la base principale. En raison du blocage du versement d'une copie de la base principale du SNDS sur les serveurs du HDH (cf. parties 2.1 et 2.2.1), la CNAM reste la seule à assurer ce stockage mais a bénéficié, pour la mise à disposition des données de la base principale, du concours du HDH qui a mis à sa disposition 2 ETP de *data managers* spécialement recrutés par le HDH pour aider la CNAM, ainsi que 2 ETP experts de *data scientists* internes.

Par ailleurs, beaucoup d'efforts ont déjà été consacrés par la CNAM à la sécurisation de l'accès aux données du SNDS, notamment par la **rénovation de sa chaîne de pseudonymisation** suite à la mise en demeure par la CNIL en avril 2018. À l'heure actuelle, le cadre réglementaire n'a confié qu'à la CNAM la mise en œuvre de l'appariement de bases de données avec la base principale du SNDS, et il n'est pas prévu de transférer la chaîne de pseudonymisation au HDH⁵³. Dans ces conditions, tous les appariements directs nécessitent, pour l'heure, une mobilisation de la CNAM⁵⁴.

b) Une opération de calibrage par la CNAM des données demandées particulièrement chronophage

Une partie des critiques se rattachant à cette étape de mise à disposition se concentre également sur la ré-interrogation par la CNAM du périmètre des données requises par un projet. En effet, au titre du **principe de minimisation des données**, il appartient à la CNAM, en sa qualité de responsable de traitement, de cibler exactement le champ des données de la base principale du SNDS nécessaires à la réalisation du projet. La finalisation du périmètre des données demandées constitue l'objet principal et sans doute l'étape la plus chronophage du conventionnement avec

⁵³ Ce dernier ne dispose pas, en tout état de cause, d'une autorisation d'accès au NIR (il lui est ainsi spécifiquement interdit d'avoir accès au NIR et aux noms et prénoms des personnes pour l'enrichissement de la base principale et la constitution des bases du catalogue, en application du dernier alinéa du III de l'article R. 1461-3 du CSP) et n'est autorisé qu'à réaliser des appariements entre les bases de données du catalogue et la base principale du SNDS (en application du 1^{er} du III de l'article R. 1461 3 du CSP).

⁵⁴ Le IV de l'article R. 1461-3 du CSP prévoit néanmoins désormais la possibilité pour la CNAV et l'INSEE de contribuer en qualité de sous-traitants pour la réalisation d'opérations d'appariement nécessaires à la constitution de bases dans le cadre du SNDS.

le porteur de projet : elle peut impliquer plusieurs itérations entre la CNAM et ce dernier pour s'entendre sur les variables qui seront effectivement mises à disposition.

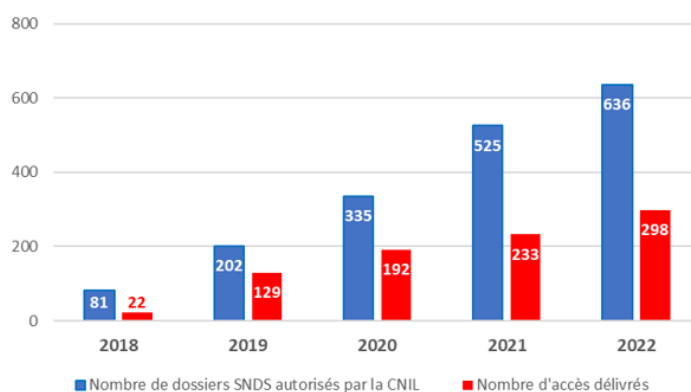
Cette étape ne manque pas d'irriter une partie des porteurs de projet qui ont le sentiment que le périmètre des données sollicitées a déjà suffisamment été interrogé au stade de l'expression des besoins devant le CESREES et la CNIL et comprennent difficilement que leur choix de variables puisse à nouveau être discuté au moment du ciblage par la CNAM. Pour d'autres, la finalisation de l'expression de leurs besoins au stade de la phase règlementaire est considérée comme prématurée, puisqu'elle suppose un accès préalable à la base requêtée pour identifier au mieux les variables et la population à cibler.

Afin de fluidifier l'étape de finalisation du périmètre des données à extraire, le HDH a **conclu une convention avec la CNAM de co-responsabilité de traitement**, permettant au HDH de gérer, par délégation, les extractions et le conventionnement pour les projets hébergés sur sa plateforme.

En définitive, l'impossibilité actuelle de transférer une copie de la base principale du SNDS sur la plateforme du HDH limite, pour l'instant, les marges de manœuvre de ce dernier pour agir efficacement en faveur d'une réduction des délais de mise à disposition des données de la base principale du SNDS.

Compte tenu du goulot d'étranglement dans la gestion des accès aux données de la base principale du SNDS au niveau de la CNAM, l'écart tend à se creuser, en cumulé, entre le nombre de demandes d'accès au SNDS autorisées par la CNIL et le nombre d'accès effectivement mis en œuvre, les seconds ne représentant que 47 % des premiers en juillet 2022, même si ces accès effectifs sont, en valeur absolue, en augmentation :

Graphique 3 : Nombre de dossiers SNDS autorisés par la CNIL et nombre d'accès délivrés par la CNAM, pour la période 2018-2022*



* Pour 2022, les données vont jusqu'en juillet.

Source : Health Data Hub.

Cette situation entame la confiance des acteurs de la recherche, notamment privée, dans la capacité de la CNAM et du HDH à assurer une mise à disposition des données du SNDS dans des temps utiles. Certains acteurs, en particulier privés, font valoir une diminution du nombre de demandes d'accès au SNDS historique. Selon ces acteurs, les industriels se détourneraient en partie du SNDS français pour privilégier des bases de données étrangères moins riches mais plus faciles d'accès. À l'issue d'une courte enquête interne, le Leem relève ainsi que, pour le seul SNDS,

42 % de ses adhérents déclarent « *renoncer souvent à des objectifs de recherche en raison des délais d'accès qui les conduisent à des pertes d'arbitrage et les amènent à restreindre les usages aux seuls cas obligatoires* ».

Le HDH confirme, pour sa part, que, si la part de projets mobilisant des données de la base principale du SNDS portés par des acteurs privés est restée stable entre 2019 et 2022 (4 projets sur 10), une baisse de moitié du nombre de projets portés par des responsables de traitement privés a bien été observée entre 2020 et 2022 (46 projets en 2022, contre 89 en 2020), la baisse des projets portés par des acteurs publics s'expliquant, elle, en grande partie par l'ouverture des accès permanents à la base principale du SNDS. Le HDH rapporte ainsi que « *différents bureaux d'études ont indiqué qu'en raison des délais d'accès importants et de l'absence de visibilité sur le calendrier de mise à disposition effectif des données, un certain nombre d'industriels de santé se détournent de cette base pour solliciter des sources de données notamment étrangères plus faciles d'accès.* »

Par comparaison, les délais d'accès aux ressources de la UK Biobank ont été divisés par plus de 6 sur la période 2012-2022.

Graphique 4 : Nombre moyen de semaines entre le dépôt de demande d'accès et l'extraction des données pour la UK Biobank



Source : UK Biobank (<https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>).

Enfin, la perspective d'une **remise en cause du dispositif des accès permanents au SNDS** par le projet de règlement sur l'espace européen des données de santé (« *European Health Data Space* » – EHDS) inquiète les acteurs qui en bénéficient (cf annexe 2). Pour mémoire, les équipes des organismes bénéficiant aujourd'hui d'un accès permanent ont la possibilité d'interroger quotidiennement la base principale du SNDS, notamment pour réaliser des **études de faisabilité**⁵⁵.

⁵⁵ En application de l'article R. 1461-17 du CSP ; une déclaration de la liste et des caractéristiques des études mises en œuvre dans le cadre de cet accès permanent doit être réalisée au HDH, mais cette déclaration est en pratique rarement mise en œuvre.

1.3 Un environnement institutionnel trop peu coopératif et un principe d'ouverture des données encore à asseoir

1.3.1 La Plateforme des données de santé (Health Data Hub) : de premières réalisations encourageantes

L'analyse des principales missions et réalisations du HDH en annexe 4 montre que ce dernier a su s'imposer en seulement quelques années comme un partenaire central de l'écosystème de la recherche sur les données de santé. Fort d'une fonction de **point d'entrée unique** des demandes de traitement de données de santé n'impliquant pas la personne humaine, le HDH est salué pour la **qualité de l'accompagnement** « de bout en bout » qu'il met à la disposition des 97 projets sélectionnés dans le cadre de ses appels à projets – dont 50 % utilisent de l'intelligence artificielle et plus de 30 % incluent des partenariats industriels –. Son expertise est désormais largement mise à profit de la constitution des entrepôts de données de santé hospitaliers et d'un entrepôt de données de médecine de ville, dans le cadre d'un appel à projets de France 2030.

En revanche, sa mission centrale de réunion et de mise à disposition des données de la base principale du SNDS a dès le départ été compromise par l'impossibilité de transférer une copie de ce dernier sur sa plateforme en raison du choix de la solution logicielle Azure de Microsoft. Les marges de manœuvre du HDH pour accélérer la mise à disposition des données de la base principale du SNDS se sont donc révélées réduites, cantonnées essentiellement, en amont, à une préparation des dossiers aux étapes réglementaires devant le CESREES et la CNIL et, en aval, à un accompagnement dans le conventionnement avec la CNAM pour l'extraction des données. La qualité de cet accompagnement technique et juridique reste toutefois appréciée par les acteurs concernés.

Sont également à souligner les efforts conduits par le HDH dans l'organisation de **formations** et d'**ateliers de sensibilisation à l'utilisation des données de santé**, notamment à l'exploitation de la base principale du SNDS, participant d'une montée en compétences de la communauté de recherche mais aussi des représentants des usagers.

Le HDH s'est en outre fortement impliqué sur la **scène européenne** dans la construction d'un espace européen des données de santé. Identifié comme référence au niveau européen, la Commission européenne l'a sélectionné pour coordonner le projet HealthData@EU en vue de construire une première version d'un espace européen des données de santé. Il a également été chargé en octobre 2022 de porter un⁵⁶ des cinq cas d'usage transfrontaliers sélectionnés par la Commission européenne pour évaluer la faisabilité de l'utilisation secondaire de données de santé de plusieurs pays européens. Enfin, le HDH s'est positionné sur plusieurs appels à projets européens (maladie rares, *data quality*, données de vie réelle...) et est lauréat d'un appel d'offres de l'Agence européenne du médicament.

Certes, les acteurs du monde de la recherche sont en attente d'un **enrichissement de la base principale du SNDS** ainsi que d'une **montée en visibilité des bases de données existantes**. Toutefois, il convient de souligner que, si aux termes de l'article R. 1461-3 du code de la santé

⁵⁶ Ce cas d'usage consiste à anticiper les parcours de soins dans les maladies cardio-métaboliques par le recours à l'intelligence artificielle

publique (1° du III), le HDH assume une responsabilité opérationnelle dans l'alimentation de la base principale du SNDS par d'autres sources de données, notamment cliniques, sociales et médico-sociales, **l'identification des composantes futures de la base principale relève du comité stratégique des données de santé**⁵⁷.

1.3.2 Une perspective de centralisation des données qui suscite des crispations importantes, et le maintien de réflexes « propriétaires »

1.3.2.1 Une notion de « catalogue » du SNDS dénaturée par le décret du 29 juin 2021 relatif au traitement de données à caractère personnel dénommé « système national des données de santé »

La constitution par le HDH du « catalogue » du SNDS, c'est-à-dire la collecte de bases de données ne couvrant pas l'ensemble de la population pour leur mise à disposition sur sa plateforme, a alimenté, chez les acteurs de la collecte de données de santé, la **Crainte d'un mouvement de centralisation** de l'ensemble des bases de données de santé sur une unique plateforme sous l'égide d'un seul opérateur.

L'élargissement du SNDS à des données cliniques est un objectif inscrit par le législateur dans la loi « OTSS » du 24 juillet 2019 qui a complété le champ des bases de données rassemblées et mises à disposition au sein du SNDS pour y intégrer les données mentionnées aux 6° à 11° de l'article L. 1461-1 du CSP, dont potentiellement toutes celles collectées à l'occasion de soins pris en charge par l'assurance maladie tant en médecine de ville qu'à l'hôpital⁵⁸. À cet égard, l'étude d'impact annexée au projet de loi « OTSS » relève que *« compte tenu de l'élargissement du périmètre du système national des données de santé, il n'est plus envisageable que la caisse nationale de l'assurance maladie, qui a la charge de la maîtrise d'ouvrage de l'applicatif, demeure l'unique responsable du traitement et centralisateur. Plusieurs responsables des traitements seront donc en charge du rassemblement et de la mise à disposition de ces données »*.

Dans son rapport⁵⁹ sur le projet de loi « OTSS », M. Alain Milon, rapporteur de la commission des affaires sociales du Sénat, y voyait un objectif de **« gestion décentralisée favorisant la création d'entrepôts de données multiples »**. Toutefois, comme l'explique M. Milon, afin de *« limiter la constitution d'entrepôts de données « secondaires » et spécifiques à certains fournisseurs de données (notamment les établissements de santé) »*, la loi « OTSS » a précisé, à l'article L. 1461-1 du CSP, que *« les responsables ou les catégories de responsables des traitements du système national des données de santé et leurs rôles respectifs sont définis par décret en Conseil d'État, pris après avis de la Commission nationale de l'informatique et des libertés. Les responsables de ces traitements sont nominativement désignés par arrêté »*.

⁵⁷ En application de l'article 1^{er} de l'arrêté du 29 juin 2021 portant création du comité stratégique des données de santé, ce dernier est notamment chargé d'*« identifier les bases de données existantes qui ont vocation à être inscrites dans l'arrêté prévu au I de l'article R. 1461-2 du code de la santé publique, recommander leur inscription et, le cas échéant, émettre un avis sur l'ordre de priorité de leur inscription »*.

⁵⁸ Cf. 6° du I de l'article L. 1461-1 du code de la santé publique.

⁵⁹ Rapport n° 524 (2018-2019) de M. Alain Milon, fait au nom de la commission des affaires sociales du Sénat, déposé le 22 mai 2019.

Il n'en demeure pas moins que la loi « OTSS », en élargissant potentiellement le champ du SNDS à l'ensemble des sources de données cliniques collectées à l'hôpital et en ville, a pu, dès le départ, semer une forme de confusion en suggérant que ces bases, par nature constituées de façon décentralisée, devaient alimenter une base, le SNDS, qui, dans l'esprit collectif, est centralisée.

À cela, s'ajoute le fait que le décret n° 2021-848 du 29 juin 2021, relatif au SNDS, pris pour l'application de la loi « OTSS », a aussi contribué remettre en question l'idée d'une organisation décentralisée des bases de données appelées à nourrir le SNDS, en introduisant une **confusion dans la notion même de « catalogue » du SNDS**.

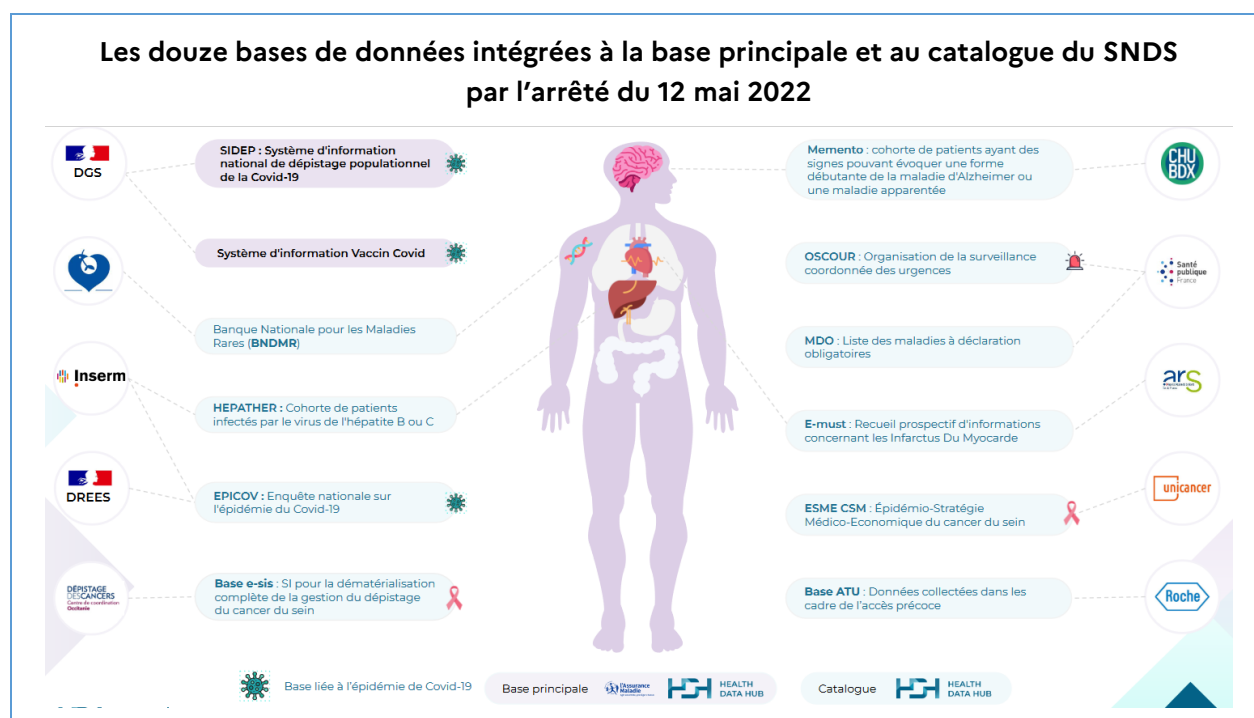
Le catalogue de données que devait mettre en place le HDH était initialement conçu par la mission de préfiguration du Health Data Hub comme un répertoire appelé à être « *enrichi progressivement pour référencer à terme les principaux gisements de données de santé financées par la solidarité nationale* »⁶⁰ et devant intégrer des jeux de données documentés par les producteurs des données et accompagnés d'un échantillon⁶¹.

Néanmoins, l'article R. 1461-2 du CSP, dans sa rédaction résultant du décret du 29 juin 2021 précité, définit le SNDS comme comprenant, d'une part, une base principale couvrant l'ensemble de la population, et, d'autre part, un ensemble de bases de données ne couvrant pas l'ensemble de la population dénommé « catalogue ». L'article R. 1461-3 du même code précise que le HDH est responsable notamment « *du stockage et de la mise à disposition des données de la base principale et de l'ensemble des bases de données du catalogue* » (2° du III) et qu'il « *détient une copie de la base principale détenue par la Caisse nationale de l'assurance maladie, ainsi qu'une copie des bases inscrites dans le catalogue* » (avant-dernier alinéa du III).

Il résulte de ces dispositions que le « catalogue » du SNDS, dont les bases constitutives sont désignées par arrêté du ministre chargé de la santé après avis de la CNIL, a été conçu par le pouvoir réglementaire non pas comme un répertoire renseignant les chercheurs sur le contenu et la gouvernance de bases de données complémentaires à la base principale du SNDS, mais comme la **réunion sur une plateforme centralisée par le HDH de copies de bases de données**.

⁶⁰ Marc Cuggia, Dominique Polton et Gilles Wainrib, *Health Data Hub*, mission de préfiguration rapportée par Stéphanie Combes, remis le 12 octobre 2018 (p. 28).

⁶¹ *Ibidem*, p. 31.



Source : Health Data Hub.

1.3.2.2 Une brèche dans la confiance avec les producteurs de données, qui n'a pas favorisé l'ouverture et le partage des données

a) Une démarche territoriale de constitution d'entrepôts de données de santé qui s'accommode mal d'un transfert de copies de leurs bases sur une plateforme nationale

Cette perspective de centralisation de l'hébergement de copies de bases de données au sein du « catalogue » du HDH a été, à tort ou à raison, mal accueillie par une large part des producteurs de bases de données entendus par la mission, que celles-ci aient un caractère local/régional (entrepôts de données de santé hospitaliers) ou national (cohortes et registres).

D'une part, certains questionnent la pertinence d'un hébergement centralisé de copie de ces bases sur une plateforme nationale, avec un risque de **couper le lien entre la donnée et son producteur** : l'impossibilité de s'appuyer sur l'expertise du producteur des données pénalise potentiellement leur exploitation et peut, par exemple, donner lieu à des biais d'analyse. L'impact écologique du stockage sur différentes plateformes de larges volumes de données est également souligné.

D'autre part, beaucoup pointent l'écueil d'une maximisation des **risques de cybersécurité** par la centralisation de l'hébergement de copies de ces bases sur une même plateforme, ainsi que la crainte pour les producteurs de données de **perdre la confiance des patients** si toutes leurs données sont centralisées en un seul et même endroit.

La pertinence d'une mise à disposition centralisée par le HDH de **grandes collections nationales de données**, telles que la banque nationale des maladies rares, la base « Maladies à déclaration obligatoire » ou encore la cohorte « MEMENTO » sur la maladie d'Alzheimer est moins discutée, notamment eu égard à l'intérêt de leur croisement avec les données de la base principale du

SNDS. Mais au total, de nombreux producteurs de données ont fait part à la mission de leur opposition à une logique centralisatrice qui consisterait à transférer systématiquement une copie des bases de données cliniques produites localement sur la plateforme du HDH.

Toutefois, certains acteurs considèrent que cette centralisation favorise le partage car toutes les bases de données ne sont pas financées pour organiser le partage de leurs données (accès distant et sécurisé, stockage...).

Dans ce contexte, une majorité des personnes entendues par la mission plaide pour un retour à une logique décentralisée dans l'organisation et l'animation des bases de données cliniques appelées à renforcer le volet médicalisé de la base principale du SNDS, **sans transfert systématique de copies de ces bases sur la plateforme du HDH**, et pour la constitution d'un catalogue au vrai sens du terme, qui serve de référencement de ces bases, notamment pour leur permettre de gagner en visibilité, avec une description claire du contenu et de la structuration de leurs données.

b) La persistance de « réflexes propriétaires »

Si l'ensemble des producteurs de bases de données interrogées par la mission affirment adhérer au principe d'ouverture et de partage des données qu'ils détiennent au service de la recherche en santé, la mission constate que ce principe coexiste avec une réalité plus nuancée, qui va parfois jusqu'à une réticence dans les faits à l'ouverture et au partage de données.

Un des éléments le plus mis en avant est le rappel que les bases de données ont été constituées par des équipes souvent au prix d'**efforts considérables**, notamment pour obtenir des financements (par exemple s'agissant des cohortes, cf. partie 1.2.3.2). Compte tenu des difficultés rencontrées dans la constitution de ces entrepôts ou cohortes, les producteurs de données restent attachés au « retour sur investissement » de ces efforts. Ils entendent en quelque sorte pouvoir « garder la main » sur la définition des conditions de la valorisation scientifique et économique de leurs bases de données.

Certains, craignant probablement une perte en termes d'opportunités de publications et, par suite, de conséquences potentielles sur les crédits liés aux SIGAPS⁶², vont jusqu'à réclamer une forme d'embargo, qui consisterait à réserver temporairement l'exploitation des données de leurs bases à leurs équipes internes, préalablement à l'ouverture de ces bases à d'autres acteurs. La Fédération des hôpitaux de France appelle ainsi à cet égard à « *protéger l'intérêt des établissements de santé publics qui doivent être les premiers bénéficiaires de leurs données* ».

D'autres soulignent la nécessité de pouvoir déterminer les conditions de valorisation économique des bases de données qu'ils constituent, afin d'amortir les investissements consentis. L'AFM-Téléthon rappelle ainsi avoir ainsi investi dans la constitution de registres cliniques dédiés à certaines maladies neuromusculaires. Il est dès lors important pour l'AFM d'être en capacité de définir les conditions d'un partage de ces coûts d'investissement avec les bénéficiaires de

⁶² Le Système d'interrogation, de gestion et d'analyse des publications scientifiques (SIGAPS) est un outil pour recenser et analyser les publications scientifiques issues des établissements de santé. Il permet d'évaluer la production scientifique des établissements de santé et de déterminer ainsi un indicateur, qui conditionne en partie le montant de l'enveloppe budgétaire allouée à la recherche (Mission d'enseignement, de recherche, de référence et d'innovation – MERRI).

l'utilisation de ces registres, dont en particulier les laboratoires pharmaceutiques, capacité dont elle estime qu'elle lui échapperait en cas de mise au catalogue du HDH de ces registres.

Signe que la culture du partage des données nécessite un temps d'appropriation, l'AP-HP indique que, pour l'heure, tout projet de recherche requérant un accès aux données de son entrepôt doit être réalisé en partenariat avec un médecin de l'AP-HP, lequel sera d'ailleurs porteur de l'instruction de la demande devant le comité scientifique et éthique de l'institution, cet élément ayant été, à l'époque de la création de l'entrepôt de données de santé de l'AP-HP, été indispensable pour fédérer la communauté hospitalière autour du projet.

La mission peut entendre voire comprendre ces réticences au partage, qui sont tant le fait des acteurs « de terrain » (cliniciens, chercheurs) que des institutions qui mettent en place et exploitent les infrastructures. Mais la mission ne peut que constater les conséquences très négatives de tels obstacles au partage : en réalité, la situation est « perdante - perdante ». Perdante pour chacun des acteurs ou des chercheurs, qui se privent des opportunités créées par l'accès facilité à d'autres bases des données. Perdante globalement pour la science, la recherche et *in fine* l'amélioration des soins, car la limitation du partage des données génère moins de recherche et moins de création de valeur scientifique, avant même la création de valeur économique. De telles réticences apparaissent en tout état de cause difficilement tenables dans la perspective de la mise en œuvre du projet de règlement EHDS dont l'article 33 consacre **l'obligation pour tout détenteur de données électroniques de les mettre à disposition à des fins d'utilisation secondaire**.

1.3.3 Des financements publics importants dans le cadre du plan d'investissement France 2030, mais une mauvaise articulation de certaines initiatives

La mission a recensé les principaux dispositifs de financements soutenant la réutilisation secondaire des données de santé (en dehors des financements des registres et des cohortes), présentés dans le tableau ci-après.

Tableau 4 : Principaux financements publics contribuant à la réutilisation secondaire des données de santé en 2023 (hors registres et cohortes)

Programme, appel à projet ou structure financée	Source du financement	Montant	Pour les appels à projet			
			Opérateurs/ pilotes	Objectif	Cible	Durée
Dans le cadre du programme Innovation France 2030						
Programme et équipement prioritaire de recherche Santé numérique (PEPR SN)	Plan Innovation santé 2030/ Stratégie Santé numérique	60 M€	DNS, ANR Pilotage scientifique : Inserm et INRIA	Financement de recherche fondamentale sur des modèles, outils et méthodes pour produire, gérer et exploiter les volumes croissants de données de santé hétérogènes	17 projets ciblés, notamment sur les maladies cardiovasculaires et neurologiques	Lancement en juin 2023 pour une durée de 7 ans
Appel à projets prématuration-maturation	Plan Innovation santé 2030/ Stratégie Santé numérique	275 M€ - dont 12M€ de la SASN	DNS, ANR	Accélérer le transfert de la recherche vers le monde socio-économique	Offices de transfert de technologies (OTT)- Financement de 17 consortiums dont un consortium lauréat en santé numérique (Inserm Transfert, SATT Sud-Est)	
AMI Santé numérique	Plan Innovation santé 2030/ Stratégie Santé numérique	55,4 M€	DNS, Bpi France	Aide à l'innovation d'entreprises ; parmi les axes de l’AMI figurait la collecte et/ou la structuration des données de santé	14 lauréats dont des entreprises spécialisées dans la donnée (notamment en appui des entrepôts de données de santé)	3 ans maximum
Accompagnement et soutien à la constitution d'entrepôts de données de santé hospitaliers	Plan Innovation santé 2030/ Stratégie Santé numérique et ONDAM	75 M€ (35M€ France 2030 et 40M€ d’ONDAM)	DGOS, DNS, BpiFrance, Support : Health data hub	Financement pour l'accompagnement et le soutien à la création d'entrepôts de données de santé (logique de co-financement)	Etablissements de santé ou groupements de coopération sanitaire - 6 lauréats annoncés, un 2ème AAP cloturé	Lancement en 2022 pour 40 mois maximum
Accompagnement de projets de recherche	Plan Innovation santé 2030/ Stratégie Santé numérique	16,1M€	ANR, IRESP, INSERM	Fresh : constituer un référencement des recueils individuels de données en santé France Cohortes : financer l’accompagnement à la création et au suivi de cohortes	Financement des projets FRESH (3,3 M€) et France Cohortes (12,8 M€	5 ans

Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé

Appel à projet imagerie	Plan Innovation santé 2030/ Stratégie Santé numérique	90 M€	DGE, DNS, Bpi France	Co-financement du développement de nouvelles solutions d'imagerie médicale et de logiciels de traitement de l'image	Start-ups et PME ou industriels en association avec des académiques et/ou des établissements de soins portant des projets collaboratifs en imagerie médicale	2023-2024
Appel à projets « Évaluation du bénéfice médical et / ou économique des DM numériques ou à base d'intelligence artificielle »	Plan Innovation santé 2030/ Stratégie Santé numérique, et FII Grand défi IA et Santé	91 M€ (dont 71M€ issue SASN et 20M€ FII grand défi IA et santé)	DNS, Bpi France	Financement de projets d'évaluation visant à évaluer le bénéfice des dispositifs médicaux numériques ou à base d'intelligence artificielle (entre 200 000 et 5M€). Note : la prochaine vague de cet AAP est fusionné avec l'AAP démonstrateur du plan DM (60M €)	Dispositif médical (DM), dispositif médical implantable actif (DMIA) ou dispositif médical de diagnostic in vitro (DMDIV) dont la composante numérique représente la part prépondérante de l'actif ou de la proposition de valeur	Durée de 12 à 48 mois
Data Challenge en santé et registres FSM	Fond FII Grand défi IA & Santé intégré à la Stratégie Santé numérique	11,52M€ (dont 1,91M pour registre FSM)	DNS, BPI France, HDH	Financement de projet d'intelligence artificielle pour répondre à une question clinique (jusqu'à 300 000€)	Société savante, établissement de santé, groupement de coopération ou Institut hospitalo-universitaire	9 lauréats sur la période 2021-22 2020-2025
Hors France 2030						
Plateforme des données de santé	Fonds de transformation de l'action publique et ONDAM	En 2023, 28 M€ dont 13 M€ d'ONDAM				
Financement d'infrastructure de recherche	Plan Médecine France génomique 2025	100 M€	ANR	Financement de plateformes telle que la plateforme France Génomique	Pour les infrastructures nationales en biologie et santé (INBS)	2023
DATAE	ONDAM –ES	3,5 M€ par an (reconduction)	DGOS, Health data hub	Appel à projets de recherche mobilisant des données collectées dans les entrepôts	Pour les entrepôts de données de santé référencés par la CNIL	1er appel à projet fin 2023- début 2024

Source : mission

Ce recensement n'a pas vocation à être exhaustif. A titre d'exemples, d'autres bases de données sont financées par le ministère de la santé, dans le cadre des plans de santé publique comme la banque nationale de données maladies rares). De même, la filière cancer dispose de modalités de financement particulière : la plateforme de données en cancérologie pilotée par l'INCA a été cofinancée par le FIAC et la banque publique d'investissement (BPI). Il existe également des financements européens, dans le cadre du programme Horizon ou du programme « UE par la santé » (EU4Health) : ainsi, le Health data hub est lauréat d'un financement pour préfigurer l'espace européen des données de santé.

Cette précaution prise, le tableau atteste **d'un investissement public important consenti pour favoriser l'utilisation des données de santé** (primaire ou secondaire) **depuis 2020**. En effet, d'importants financements ont été accordés dans le cadre du programme d'investissement France 2030, notamment dans le cadre de la stratégie d'accélération du numérique en santé. Dans une moindre mesure, des financements ont été mobilisés dans le cadre de l'ONDAM, notamment pour le financement d'une partie du budget du Health data hub ou pour le financement des entrepôts de données hospitaliers.

Les financements actuels comportent toutefois deux limites.

D'une part, les financements sont éclatés entre plusieurs pilotes ou opérateurs, ce qui conduit au financement de projets dont l'articulation n'est pas claire, voire qui se recoupent pour partie. Ainsi, l'articulation entre le projet France Cohortes (détail en encadré *supra*) et la plateforme Health data hub n'est pas claire : les données des cohortes existantes ont-elles vocation à intégrer l'outil France cohortes ou la plateforme des données de santé ? France cohortes a-t-il vocation à suppléer le Health data hub pour les cohortes ? Laquelle de ces structures a-t-elle vocation à favoriser la réutilisation des données des cohortes ? La nécessité d'une articulation entre les deux démarches était pourtant pointée dès 2019 par le rapport IGAS/IGESR sur les cohortes.

Un autre exemple de recoupement entre projets concerne les référencements de bases de données de santé menés conjointement par le Health data hub et l'Inserm (via le projet FRESH).

Le projet FRESH

Le projet FRESH (France - Recherche en Santé Humaine) mené par l'IRESP, financé par France 2030 (3,3 M€ sur 5 ans), consiste en un référencement de l'ensemble des recueils de données individuelles de santé (études interventionnelles et non interventionnelles), au niveau national et international.

Il s'appuiera sur l'acquis du portail Epidémiologie France, qui constitue un premier recensement des bases de données en santé françaises (en particulier les cohortes et les registres), réalisé par Aviesan. A ce stade, il est prévu qu'il « moissonne » les données du Health data hub (notamment en ce qui concerne les métadonnées) ainsi que d'autres plateformes (comme clinical trial.gov à l'échelle internationale). Il ambitionne de favoriser l'harmonisation des pratiques de recueils de données de santé.

Le financement a été obtenu à l'été 2023. Une première version étendue par rapport au Portail Epidémiologie France existant devrait être mise en ligne fin 2024. D'autres versions incluant des fonctionnalités avancées et davantage de sources sont prévues en 2025.

Dans le même temps, le Health data hub porte actuellement un projet de répertoire public de référencement des bases de données en santé, en cohérence avec ses missions de promotion de l'innovation dans l'utilisation des données de santé, par exemple par la mise en relation des acteurs, la diffusion d'outils et de contribution à l'élaboration de référentiels et de diffusion de normes de standardisation pour l'échange et l'exploitation des données de santé⁶³.

Le périmètre du projet FRESH est plus large, puisqu'il porte aussi sur les études interventionnelles. Sur ce dernier champ, il existe également un projet de base nationale des essais cliniques (projet ECLAIRE également financé dans le cadre de France 2030), qui vise à référencer les essais cliniques portée par le ministère de la santé (délégation du numérique en santé et direction générale de l'offre de soins, en associant l'agence du numérique en santé).

Si l'objectif de mieux cartographier les bases de données en santé de manière transverse est indispensable compte-tenu du manque de lisibilité actuelle, sa réalisation parallèle par plusieurs acteurs publics atteste d'un manque de coordination interministérielle.

La seconde limite de ces financements tient à leur caractère non pérenne. C'est notamment le cas du financement des entrepôts de données de santé, le principal appel à projet permettant un co-financement jusqu'à 50% des coûts de financement des entrepôts jusqu'en 2026. Si ce financement a été très largement salué, car il a permis de conforter le développement des entrepôts auparavant autofinancés par les établissements de santé, l'absence de financement pérenne est regrettée, d'autant que les coûts fixes des entrepôts ont été récemment objectivés. Sous l'égide du comité stratégique des données de santé, un groupe de travail réuni en 2023 a estimé à en moyenne à **2M€ par an le coût de développement d'un EDS sur 5 ans puis à 3M€ par an le coût de financement de son fonctionnement nominal**. Les coûts estimés dans sept établissements sont toutefois très variables selon les établissements (voir détail dans l'annexe 1).

Par ailleurs, les financements des entrepôts liés à la mise à disposition de données pour des acteurs externes restent limités. Même pour des bases de données matures, **les financements issus de la valorisation de l'exploitation des données sont très loin de couvrir l'ensemble des coûts**. A titre d'exemple, la banque nationale de données maladies rares a perçu 1,5 millions d'euros grâce à une valorisation active de ses données depuis deux ans, pour un coût complet estimé entre 30 et 40 millions d'euros depuis dix ans. La réglementation européenne pourrait également contraindre la capacité des entrepôts à valoriser leurs données (cf. partie suivante).

1.3.4 Des partenariats économiques tâtonnants, parfois décourageants et peu compétitifs, en l'absence de modèle de tarification lisible

En complément des financements publics, les acteurs qui ont structuré des bases de données ont été incités à construire des partenariats avec les acteurs privés pour la valorisation de ces bases de données. Les cohortes constituent une bonne illustration de la construction de tels partenariats publics-privés (détail dans l'encadré *infra*).

⁶³ Arrêté du 29 novembre 2019 portant approbation d'un avenant à la convention constitutive du groupement d'intérêt public « Institut national des données de santé » portant création du groupement d'intérêt public « Plateforme des données de santé »

La valorisation économique des données de santé issues de cohortes par Inserm Transfert

Inserm Transfert valorise depuis 12 ans les données de santé issues des cohortes, des essais cliniques sous promotion Inserm et des bases de données de santé (BRCA Share) au travers de partenariats innovants et ambitieux avec les acteurs industriels français et internationaux. Depuis 2012, 38 industriels sont devenus partenaires de 21 cohortes ; 44 essais cliniques ont fait l'objet de partenariats avec 22 industriels, certains industriels étant partenaires de plusieurs cohortes et/ou essais cliniques.

Ces acteurs industriels sont majoritairement des laboratoires pharmaceutiques, des entreprises de taille intermédiaire ou des acteurs du diagnostic. Plus récemment, Inserm Transfert a observé un fort intérêt d'autres types de structures comme des biotech ou des startups pour réaliser des projets nouveaux notamment dans le domaine de la santé numérique.

Dans le cas des cohortes, les partenariats proposés par Inserm Transfert visent les deux objectifs suivants :

1. Rendre accessibles à la communauté industrielle les données de santé Inserm ;
2. Participer financièrement au développement et au maintien des cohortes en complétant leur budget global qui n'est jamais totalement pris en charge par les financements publics (objectifs de valorisation de cohortes labélisées investissement d'Avenir).

Deux types de partenariats sont proposés :

- Les partenariats public/multiprivé proposent aux industriels de devenir partenaires d'une cohorte sur une durée pluriannuelle (3 ans), renouvelable. Ces partenariats sont particulièrement adaptés pour des cohortes en population générale ou de patients atteints de pathologies dont les axes thématiques seraient au cœur de la stratégie de développement de plusieurs industriels qui envisageraient notamment de réaliser plusieurs projets d'utilisation secondaire des données issues d'une même cohorte. A ce jour 25 partenariats public/multiprivé ont été signés sur 4 cohortes (Hepather, CKD REIN, Cobra et Constances) pour un montant global de 26 M€ et 57 projets d'utilisation secondaire des données ont été réalisés pour un budget total de 2 M€.

- Les partenariats au projet interviennent dans le cas où l'industriel a déjà une idée assez précise des objectifs qu'il souhaite atteindre grâce à la réutilisation des données d'une cohorte. L'industriel devient partenaire de la cohorte pendant la durée de son projet d'utilisation secondaire des données. Il s'engage financièrement auprès de la cohorte, selon les spécificités de celle-ci, et prend en charge des coûts directs de réalisation de son projet. A ce jour 25 partenariats au projet ont été signés pour une vingtaine de cohortes pour un montant global de plus de 12 M€ pour la participation au développement des cohortes et de 5 M€ pour la réalisation des projets d'utilisation secondaire des données.

Source : *Inserm transfert*

Ces partenariats sont toutefois freinés par plusieurs facteurs.

Tout d'abord, les acteurs économiques soulignent que **les négociations et la contractualisation avec les acteurs publics occasionnent des délais qui peuvent être particulièrement élevés, en particulier du fait d'un manque de visibilité sur les modalités de tarification**. Plusieurs acteurs ont

souligné que le manque de visibilité en amont sur le contenu des bases de données disponibles et leur qualité ainsi que sur les tarifs applicables est source d'une grande incertitude et constitue une des raisons de renoncements *a posteriori* à des objectifs ou à des projets de recherche.

Ainsi, les conclusions du groupe de travail portant sur la tarification des données des entrepôts de données de santé, dans le cadre du comité stratégique des données de santé, ont souligné « l'absence d'un modèle ou d'un outil de tarification standardisé [qui] suppose d'itérer à plusieurs reprises au sein des équipes de l'organisme dépositaire ainsi qu'avec les équipes du porteur de projet pour stabiliser les modalités de cette tarification et pour choisir le cadre contractuel qui correspond à la demande (par exemple prestation versus collaboration). Plusieurs utilisateurs ont ainsi mentionné une attente supérieure à un an entre le dépôt de leur demande et l'accès aux données afin de s'accorder sur les contrats et sur les montants des redevances. »

L'absence d'un modèle de tarification unifié à l'échelle nationale provoque une **forte variabilité de coûts d'accès pour les utilisateurs entre les différents dépositaires**. Selon les utilisateurs interrogés par le groupe de travail précité, les devis fournis pour un même projet peuvent être multipliés par dix en fonction des porteurs de projets.

Ainsi, certaines négociations échouent du fait d'exigences perçues comme disproportionnées par les acteurs économiques, soit du fait de montants financiers jugés prohibitifs et peu compétitifs par rapport à d'autres pays étrangers, soit du fait de conditions juridiques jugées inacceptables. Ainsi, plusieurs établissements publics recherchent un partage de propriété intellectuelle très souvent bloquant pour les start-ups, en raison de réticences fortes de leurs investisseurs. D'autres établissements conditionnent les partenariats à la participation d'un de leurs chercheurs aux projets, ce qui n'est pas toujours souhaité par les entreprises.

Les négociations de partage de valeur peuvent être longues et complexes y compris entre acteurs publics, comme le souligne l'exemple de l'INCA, pour la plateforme en cancérologie, pour laquelle les négociations avec les registres de cancérologie ont été très ardues.

Enfin, même s'il existe un potentiel de valorisation économique encore inexploité, les acteurs sont unanimes pour souligner que **cette valorisation ne pourra couvrir l'intégralité des budgets de fonctionnement des bases de données**. D'un point de vue juridique, la capacité des producteurs de données à valoriser leurs données est contrainte par le cadre européen. Le règlement européen sur la gouvernance européenne des données, entré en vigueur en juin 2022 et applicable depuis septembre 2023, encadre les redevances pouvant être perçues par des producteurs de données, qui doivent être fondées sur les coûts liés au traitement des demandes de réutilisation des données⁶⁴ (article 6). Cette approche d'une tarification au coût marginal va à l'encontre d'une tarification fondée sur les coûts complets. Toutefois, aucune jurisprudence n'est encore venue préciser la portée de cette définition. De plus, les négociations en cours sur le règlement d'espace européen des données de santé pourraient conduire à des ajustements de cette règle pour les données de santé.

⁶⁴ Coûts de reproduction, de fourniture et de diffusion des données ; coûts liés à l'anonymisation ; coûts liés à la maintenance, à l'assistance aux réutilisateurs, etc.

Au-delà de ces questions juridiques, les exemples rencontrés par la mission montrent que même pour des bases matures, avec des données qualifiées, comme la banque nationale de données maladies rares, certaines cohortes ou certains entrepôts de données de santé, la valorisation économique ne permet de couvrir qu'au plus 30% du coût des bases. **Dans la plupart des cas, les redevances perçues ne couvrent et ne couvriront qu'une part minoritaire du coût des bases de données (constitution et exploitation).**

1.3.5 Un consentement des citoyens complexe

En matière de réutilisation des données de santé à des fins de recherche, prévaut, conformément à l'article 74 de la loi « Informatique et libertés », un **régime de non-opposition**, ne nécessitant pas la collecte systématique d'un consentement exprès, contrebalancé par une **obligation d'information individuelle**.

En effet, en vertu de l'article 12 du RGPD, le responsable d'un traitement de données à caractère personnel est tenu de fournir aux personnes concernées une information présentée « *d'une façon concise, transparente, compréhensible et aisément accessible, en des termes clairs et simples* », dans le cadre aussi bien d'une étude prospective impliquant la collecte des données directement auprès de la personne (article 13 du RGPD) que d'une étude rétrospective impliquant la réutilisation des données d'une base existante (article 14 du RGPD). L'article 14 du RGPD précise ainsi les modalités de cette information dans le cadre d'une utilisation secondaire, lesquelles sont applicables aux projets de recherche n'impliquant pas la personne humaine.

Le contenu d'une lettre d'information pour les études rétrospectives (utilisation secondaire)

- ✓ L'**identité** et les **coordonnées** du responsable de traitement (nom, prénom, structure, email, adresse postale...);
- ✓ La **finalité** du traitement de données (présentation synthétique et claire du projet de recherche);
- ✓ La **base juridique** du traitement au sens des articles 6 et 9 du RGPD (par exemple obligation légale, exécution d'une mission d'intérêt public, consentement, intérêts légitimes du responsable de traitement);
- ✓ La **durée de conservation** des données à caractère personnel ou les **critères utilisés** pour déterminer cette durée;
- ✓ Les **destinataires** ou les catégories de destinataires des données (les personnes qui auront accès aux données dans le cadre de la recherche);
- ✓ Le **droit de demander** l'accès aux données, la rectification ou l'effacement de celles-ci ou la limitation du traitement relatif à la personne concernée et le droit de s'opposer à l'utilisation des données;
- ✓ Les **modalités d'exercice** de ces droits;
- ✓ Le **caractère facultatif de la participation** de la personne à cette étude (les personnes sont informées qu'il n'y a pas d'obligation de participer, qu'elles peuvent refuser que ses données soient utilisées sans conséquences particulières);
- ✓ Le cas échéant, le **transfert de données** à caractère personnel hors de l'Union européenne et la référence aux garanties appropriées et les moyens d'en obtenir une copie ou l'endroit où elles ont été mises à disposition;
- ✓ Le cas échéant, [l'existence d'une prise de décision automatisée, y compris un profilage](#);
- ✓ Le droit d'[introduire une réclamation auprès de la Commission Nationale de l'Informatique et des Libertés \(CNIL\)](#);
- ✓ Les **coordonnées** (email et postale) du **délégué à la protection des données (DPD)** du responsable de traitement.

Source : Health Data Hub (<https://www.health-data-hub.fr/information-patient>).

Cette obligation d'information individuelle des personnes dont les données sont soit recueillies, soit transmises, est reprise à l'article 69 de la loi « Informatique et libertés » : **dès lors qu'il est en capacité de contacter les personnes dont les données seront réutilisées, l'auteur d'une étude rétrospective est tenu⁶⁵ de s'acquitter de cette obligation d'information individuelle.**

⁶⁵ Le non-respect par le responsable du traitement de ses obligations en matière d'information individuelle des patients l'expose à une sanction pénale pouvant aller jusqu'à 5 ans d'emprisonnement et 300 000 euros d'amende, en application de l'article 226-16 du code pénal.

En vertu du paragraphe 5, point *b* de l'article 14 du RGPD, le responsable de traitement peut néanmoins, lorsque « *la fourniture de telles informations se révèle impossible ou exigerait des efforts disproportionnés* »⁶⁶ ou lorsque l'obligation d'information « *est susceptible de rendre impossible ou de compromettre gravement la réalisation des objectifs* » du traitement, demander à la CNIL d'y déroger, sous réserve de mettre en place des garanties destinées à compenser l'absence d'information individuelle.

Selon les informations transmises par la CNIL, lorsqu'est en jeu l'information des personnes concernées quant à la constitution d'une base ou d'un entrepôt à partir de leurs données de santé, ces garanties doivent, *a minima*, assurer une information collective, qui peut prendre la forme d'une publication sur le site Internet du responsable, les réseaux sociaux, les médias régionaux, ou *via* les associations de patients ou des communiqués de presse, ou encore *via* un affichage dans les centres participants.

Lorsqu'est en jeu la réutilisation de leurs données déjà versées dans une base ou un entrepôt, un **portail de transparence** peut être mis en place afin de faciliter l'information des personnes concernées, en recensant l'ensemble des projets menés à partir des données de la base ou de l'entrepôt. **Lorsque le responsable de traitement sollicite une dérogation à l'obligation d'information individuelle** sur le fondement de l'article 14, paragraphe 5, point *b* du RGPD, **il ne peut se prévaloir de la conformité à la MR-004⁶⁷ et doit nécessairement, pour bénéficier d'une telle dérogation, obtenir une autorisation de la CNIL.** En revanche, les MR-005, 007 et 008, qui portent sur les traitements de données provenant exclusivement du SNDS, permettent la mise en œuvre de traitements de données sans information individuelle des personnes concernées.

S'agissant de la réutilisation de données d'un entrepôt de données, les personnes sont considérées comme ayant bien été informées individuellement de la réutilisation de leurs données, conformément à la MR-004, à la condition qu'elles aient été informées individuellement du versement de leurs données dans l'entrepôt et aient été renvoyées vers un portail de transparence qui détaille l'ensemble des projets de recherche réalisés à partir de l'entrepôt. Par ailleurs, il est également possible de solliciter une décision unique pour la mise en œuvre de traitements pour lesquels il n'est pas prévu d'information individuelle des personnes.

D'une façon générale, **la fourniture d'une information individuelle dans le cadre d'une étude prospective est rarement possible en raison de sa complexité et de son coût**, si bien que les porteurs de projets invoquent très souvent l'exception de l'article 14, paragraphe 5, point *b* du RGPD. Plusieurs acteurs regrettent que l'obligation d'information individuelle pour permettre à la personne concernée d'exercer son droit d'opposition doive intervenir tant à l'étape de la constitution de la base de données qu'à celle de la réutilisation des données de ces données à l'occasion d'un projet de recherche. C'est par exemple le cas pour les cohortes dont l'information initiale à l'occasion de leur constitution ne prévoyait pas l'appariement avec d'autres bases de données comme le SNDS, lequel appariement requiert alors une nouvelle information des patients concernés, retardant la réalisation d'études avec appariement à partir de ces cohortes.

⁶⁶ Ce qui est notamment le cas lorsque les données sont trop anciennes et que l'identité des personnes concernées ne peut être retrouvée, lorsque le nombre des personnes concernées est trop important ou encore lorsque le coût et le temps de la délivrance de l'information est trop important.

⁶⁷ Qui prévoit que les personnes concernées par la recherche ou leurs représentants légaux sont préalablement et individuellement informés de la réutilisation de leurs données.

Par ailleurs, il convient de rappeler que l'article 21 du RGPD, paragraphe 6, ménage la possibilité d'exclure l'exercice par la personne concernée de son droit d'opposition à l'utilisation de ses données à caractère personnel à des fins de recherche scientifique ou historique ou à des fins statistiques en application de l'article 89, paragraphe 1, **dès lors que le traitement est nécessaire à l'exécution d'une mission d'intérêt public.**

Si le cadre réglementaire posé par le RGPD offre déjà une certaine souplesse dans la mise en œuvre des obligations d'information des patients dont les données de santé sont utilisées à des fins de recherche, un travail de pédagogie auprès des porteurs de projets, dans lequel s'implique déjà le HDH, s'impose pour clarifier les conditions de respect de ces obligations et garantir que la question de l'information des patients, indispensable pour l'exercice de leurs droits, ne pèse pas excessivement dans la mise en œuvre des projets.

Au-delà, une meilleure information sur les résultats de recherche est attendue par les patients qui consentent à la réutilisation des données de santé qui les concernent. Cet effort de transparence serait de nature à renforcer l'adhésion des citoyens pour la réutilisation des données de santé.

2 Décloisonner les données de santé et fédérer l'écosystème

2.1 Un préalable : instaurer davantage de confiance en arbitrants la question de l'hébergement de la copie de la base principale du SNDS par le Health data hub (HDH)

Faire de la confiance un principe cardinal

La mission souhaite, avant toute recommandation opérationnelle, formuler une orientation générale qui doit être tout sauf un vœu pieux : que la confiance soit le maître mot de l'utilisation secondaire des données de santé. Certes, un niveau élevé de protection est absolument nécessaire, compte tenu de la sensibilité des données de santé, même pseudonymisées. Certes, chacun des acteurs a des objectifs légitimes qui peuvent, à court terme, ne pas lui sembler compatibles avec une logique de pleine ouverture et partage des données de santé à des fins d'utilisation secondaire. Mais tous ces éléments doivent être seconds eu égard à l'extraordinaire potentiel en matière de santé, de sciences et de recherche d'une pleine utilisation secondaire des données de santé. Ne pas se permettre collectivement de réaliser ce potentiel est un gâchis collectif, coûteux y compris en termes de vies qui pourraient être sauvées. C'est pourquoi la mission fait de la confiance un principe cardinal que l'ensemble des acteurs devraient toujours garder à l'esprit.

Au titre de cet impératif de confiance, la mission a décidé de traiter comme un préalable le sujet de l'hébergement de la plateforme du HDH et d'une copie de la base principale du SNDS, car elle a fait le constat que ce sujet, au-delà d'avoir un impact réel sur la capacité du HDH d'assurer pleinement sa mission et sur le formidable potentiel d'exploitation des données du SNDS, entraînait un blocage plus large et risquait de devenir le symbole d'un échec. La mission ne considère pas que ce sujet soit le sujet principal et encore moins le seul sujet : c'est pourquoi elle n'y a consacré que quelques pages et trois recommandations parmi les 37 recommandations au total. Mais c'est en revanche un sujet préalable, car le risque est que s'il n'est pas traité rapidement et clairement, les autres actions nécessaires à une meilleure et plus rapide utilisation secondaire des données de santé ne puissent être mises en œuvre ou en tout cas ne suscitent pas le plein engagement de l'ensemble des acteurs. Cette logique, que la mission assume, est aussi un double « pari » :

- pari que ce sujet préalable soit endossé, fasse l'objet de décisions rapides et avance effectivement à la hauteur des engagements et promesses, en particulier des opérateurs du cloud ;
- pari que l'ensemble des acteurs du système acceptent de s'engager franchement et sans arrière-pensées dans un plus grand partage des données de santé au service de l'intérêt collectif, dès lors que la confiance aura été rétablie par la résolution de cette problématique préalable. Mais ce double pari mérite d'être pris.

Il est donc impératif de sortir de la situation de blocage induite par le recours à la solution logicielle Azure pour l'hébergement de la plateforme technologique du HDH. Cette situation obère la capacité du HDH à assurer sa mission fondamentale de mise à disposition des données

de la base principale du SNDS, faute de pouvoir disposer en propre d'une copie de cette base principale. Cette situation altère également la confiance que doivent placer en lui les autres opérateurs et les utilisateurs de l'utilisation secondaire des données de santé, dont il a vocation à animer l'écosystème.

La mission n'entend pas revenir sur le choix de cette solution et les raisons techniques qui l'ont sous-tendu, ni sur le débat juridique⁶⁸ concernant sa conformité avec l'exigence de protection des données personnelles contre les lois extraterritoriales américaines. Elle prend acte de ce que la migration de la plateforme du HDH sur un cloud souverain a déjà été actée par le ministre des solidarités et de la santé, lequel, dans un courrier adressé à la présidente de la CNIL du 19 novembre 2020, a déclaré souscrire à l'objectif que soit adoptée une nouvelle solution technique permettant de ne pas exposer l'accès aux données du SNDS aux risques d'ingérences de pays tiers.

Cet arbitrage politique a été confirmé et a conduit à ce que la demande d'autorisation que le HDH avait déposée auprès de la CNIL pour héberger le SNDS sur sa plateforme technologique soit retirée en janvier 2022. Le HDH s'est également engagé dans l'étude d'un plan de réversibilité garantissant le transfert de sa plateforme vers une plateforme hébergée par une solution souveraine, tout en préservant un niveau exigeant de fonctionnalités, tant en termes de sécurité que de services opérationnels pour les chercheurs. Le HDH et la direction interministérielle du numérique (DINUM) ont, à ce titre, travaillé étroitement ensemble sur l'identification des services essentiels – notamment des fonctionnalités en matière de sécurité – que les fournisseurs de cloud qualifiés « SecNumCloud »⁶⁹ doivent à terme offrir pour permettre au HDH d'opérer dans des conditions optimales de sécurité et, autant que faire se peut, à l'état de l'art de la recherche en santé.

La réalisation de l'objectif d'une migration de la plateforme du HDH vers une solution de cloud de confiance, inscrite dans sa feuille de route pluriannuelle 2023-2025, est désormais fixée pour 2025.

La mission salue donc les efforts qui ont été déployés tant par le HDH que par d'autres administrations et opérateurs de l'État (DINUM, ANSSI, CNIL...) pour réfléchir aux conditions d'une migration vers un cloud souverain qui préserve les capacités de la recherche en santé, moyennant une période transitoire dont les modalités restent à déterminer.

Il s'agit désormais d'acter publiquement et officiellement cette orientation, d'en préciser les modalités et d'en tirer toutes les conséquences, en particulier s'agissant de la période transitoire le temps qu'une solution d'hébergement pleinement satisfaisante, en termes de sécurité et de souveraineté, soit effectivement disponible.

En contrepartie de ces efforts qui participent du renforcement de la confiance auprès tant de l'écosystème de la recherche en santé que des citoyens – l'engagement du transfert de la plateforme du HDH sur une solution souveraine constituant une garantie forte à cet égard –, **la mission insiste sur la nécessité pour l'ensemble des acteurs, en particulier les producteurs et**

⁶⁸ D'autant que l'annonce de la décision d'adéquation de la Commission européenne du 10 juillet 2023 approuvant le « *EU-US Data Privacy Framework* » est susceptible de modifier les termes de ce débat.

⁶⁹ Référentiel établi par l'Agence nationale de la sécurité des systèmes d'information avec un très haut niveau d'exigence en matière de sécurité informatique. Il explicite notamment des critères de protection vis-à-vis des lois extra-européennes.

détenteurs de données, de s'engager désormais pleinement dans une logique de partage effectif des données au service de la recherche et de l'amélioration des soins.

Arbitrer la question de l'hébergement de la copie de la base principale du SNDS par le Health data hub

L'écosystème du cloud de confiance est en cours de structuration et s'articule notamment autour de quatre fournisseurs ayant obtenu la qualification « SecNumCloud » établie par le référentiel de l'ANSSI : Cloud Temple, Outscale (initiative Numspot), OVHCloud et Worldline. La stratégie « cloud au centre »⁷⁰ est appelée à favoriser la montée en gamme de l'offre de cloud française, notamment eu égard à la diversité des besoins de l'État en matière d'hébergement de données sensibles. Si les acteurs français restent conscients de l'écart qui les sépare encore des « hyperscalers » américains⁷¹ en termes de richesse fonctionnelle, certains de ceux que la mission a pu rencontrer tablent sur une mise à disposition des fonctionnalités prioritaires requises par la plateforme du HDH d'ici la fin de l'année 2024.

Aux côtés de ces fournisseurs de cloud, se sont développées deux solutions « hybrides » de clouds qui se sont donné pour objectif d'obtenir la qualification SecNumCloud à horizon 18-24 mois et reposent sur un partenariat entre des hyperscalers américains et des acteurs nationaux : BLEU, constitué entre Capgemini, Orange et Microsoft, et S3NS, constitué entre Thalès et Google. Il s'agit, en quelque sorte, de la réponse des acteurs américains du cloud à la problématique de l'immunité extraterritoriale pour l'hébergement de données sensibles en France, et plus largement en Europe. L'objectif de ces partenariats est ainsi de mettre à disposition une solution de cloud qui ne soit pas soumise aux législations extra-européenne, qui puisse être certifiée SecNumCloud⁷², tout en offrant l'essentiel des services et fonctionnalités assurés par les hyperscalers américains.

Le marché du cloud souverain connaît donc assurément une dynamique, susceptible d'être confortée par l'adoption – dont l'échéance est néanmoins encore incertaine à ce stade – au niveau européen du schéma européen de certification des services de cloud (« *European Cybersecurity Certification Scheme for Cloud Services* » – EUCS).

Toutefois, la mission n'est pas en mesure d'évaluer ni de confirmer le degré de préparation des différentes solutions souveraines, françaises ou « hybrides », la crédibilité de leur calendrier d'obtention de la qualification « SecNumCloud » et leur capacité à répondre aux besoins de la plateforme du HDH. À ce stade, il ressort des travaux menés par la DINUM que l'offre qualifiée SecNumCloud n'est pas encore en capacité d'accueillir la plateforme du HDH, y compris en termes de fonctionnalités de sécurité, mais qu'au moins trois fournisseurs de cloud français se sont engagés dans une démarche de montée en gamme de leurs services pour répondre aux besoins des plateformes de données de l'État.

⁷⁰ Circulaire n° 6282-SG du 5 juillet 2021 relative à la doctrine d'utilisation de l'informatique en nuage par l'État « cloud au centre », actualisée par la circulaire n° 6404/SG du 31 mai 2023.

⁷¹ Essentiellement les « GAFAM », dont Microsoft, Google et Amazon.

⁷² Sous réserve que la participation de l'opérateur soumis à la juridiction des lois américaines ne dépasse pas 24 % du capital.

Dans le même temps, la pression s'accroît au niveau politique pour que les données sensibles des Français soient hébergées dans un environnement souverain pleinement sécurisé, comme l'ont montré les débats parlementaires sur le projet de loi visant à sécuriser et réguler l'espace numérique : tant le Sénat⁷³ que l'Assemblée nationale⁷⁴ ont œuvré pour y introduire des dispositions consacrant l'obligation pour les prestataires de services de cloud chargés de traiter des données d'une sensibilité particulière de les prémunir contre tout accès non autorisé par des autorités publiques d'États en dehors de l'Union européenne⁷⁵. Le texte adopté par l'Assemblée nationale ménage néanmoins la possibilité pour l'opérateur ayant déjà engagé un projet nécessitant le recours à un service de cloud de solliciter une dérogation à cette obligation, sans que soit connu à ce stade le résultat de l'éventuel compromis qui pourrait se dégager en commission mixte paritaire.

L'ensemble de ces éléments conduit la mission à recommander de fixer officiellement un cap, afin de donner de la visibilité à l'ensemble des acteurs. Il s'agit d'acter publiquement, à un niveau politique, l'arrêt de l'hébergement sur Azure de la plateforme du HDH, et, en parallèle de lancer les travaux pour l'hébergement du HDH, y compris d'une copie du SNDS, sur un cloud qualifié SecNumCloud. La qualification SecNumCloud, quoique particulièrement exigeante, apparaît nécessaire pour établir un environnement de confiance, avec des garanties fortes en matière de sécurité.

Pour qu'une telle programmation soit crédible, il est nécessaire de lui accoler un calendrier : les informations réunies par la mission permettent d'estimer qu'à ce stade un horizon de 24 mois est ambitieux mais crédible. La réussite d'une telle opération, dans un tel calendrier, nécessite une forte mobilisation des acteurs du cloud pour élever leur offre de services et obtenir les qualifications SecNumCloud, mais également un pilotage vigoureux de cette opération par le HDH et les acteurs publics impliqués, le cas échéant dans le cadre d'une gouvernance spécifique – associant les ministères de la santé (notamment en s'appuyant sur la DNS et la DREES) et de la recherche, mais également le ministère de l'économie et la Dinum. Ces deux derniers veilleront à la cohérence de cette opération avec la politique industrielle de l'État en matière de cloud souverain.

Recommandation n°1 Programmer l'arrêt de l'hébergement sur Azure de la plateforme du HDH et lancer les travaux pour l'hébergement du HDH, y compris d'une copie de la base principale du SNDS, sur un cloud qualifié SecNumCloud, à horizon de 24 mois, échéance ambitieuse mais crédible à ce stade.

Par ailleurs, eu égard à ce calendrier, qui doit intégrer l'opération de migration de la plateforme du HDH, la période transitoire, que la mission évalue au minimum à deux ans, ne doit pas être une période de paralysie, qui conduirait à laisser perdurer la situation actuelle de délais excessivement longs pour exploiter les données de la base principale du SNDS. Il faut donc, en parallèle de la décision de principe qui est formulée dans la recommandation n°1, arbitrer dès maintenant entre des solutions transitoires qui permettraient au HDH d'enfin piloter la gestion et la mise à disposition d'une copie du SNDS.

⁷³ https://www.senat.fr/amendements/2022-2023/778/Amdt_114.html.

⁷⁴ <https://www.assemblee-nationale.fr/dyn/16/amendements/1674/AN/1138>.

⁷⁵ Article 10 bis A du projet de loi dans sa version résultant des travaux en première lecture de l'Assemblée nationale.

Une première option consisterait à autoriser le HDH à continuer à opérer sur l'environnement Azure et d'y héberger, uniquement à titre transitoire, une copie de la base principale du SNDS, compte tenu de l'engagement ferme des pouvoirs publics et de l'opérateur de procéder à la migration de sa plateforme sur un cloud souverain à un horizon acceptable. Cette option, qui présente l'avantage de ne pas dégrader la qualité des services proposés aux porteurs de projet, paraît à la mission difficilement tenable, en particulier sur le plan politique.

L'option alternative serait de garantir, pour cette période transitoire, le transfert d'une copie de la base principale du SNDS sur une solution d'hébergement souveraine.

Cette solution souveraine pourrait être assurée par un opérateur, qui serait en capacité d'assurer dès maintenant les fonctionnalités minimales nécessaires, permettant à la plateforme du HDH de fonctionner et d'héberger une copie de la base principale du SNDS.

La mission considère qu'il faut inclure dans l'examen de cette option transitoire le centre d'accès sécurisé aux données (CASD), qui héberge d'ores et déjà une copie du PMSI depuis 2015 et a accueilli plus d'une centaine de projets impliquant des extractions de données du SNDS, notamment dans le cadre de leur appariement avec des données de cohortes (par exemple, avec la cohorte Constances). Certifié HDS et homologué au titre du référentiel SNDS, le CASD fonctionne sur le principe de la mise à disposition des porteurs de projet de boîtiers (« SD-box »), déployant une bulle sécurisée sur laquelle sont reçues les données et empêchant l'export ou l'impression de fichiers sans une validation préalable. Si le CASD ne prétend pas offrir des services équivalents aux fournisseurs de cloud américains, il permet néanmoins la réalisation de projets de recherche exploitant déjà la base principale du SNDS dans un environnement de confiance. Il serait dès lors susceptible d'opérer comme prestataire du HDH pour l'hébergement d'une copie de la base principale du SNDS.

La mission recommande donc la mise en œuvre d'un audit « flash » destiné à expertiser la capacité immédiate d'un hébergeur, cloud ou CASD, pour l'hébergement d'une copie complète du SNDS, à effectivement assurer en routine, en tant que prestataire du HDH, les extractions de la base principale et les opérations d'appariement requises par les projets de recherche en santé, et à ainsi alléger significativement la charge qui pèse sur la CNAM dans ce domaine. Cet audit devrait également évaluer dans quelle mesure l'environnement de travail offert permettrait de répondre, sur le plan technique, aux besoins des projets de recherche, eu égard aux fonctionnalités requises pour la réalisation de ces projets (langages informatiques disponibles, capacités de stockage et de calcul...).

Il ne s'agit pas ici de définir une solution pérenne – objet de la recommandation n°1 – mais une solution transitoire, mobilisable à court terme – idéalement moins de six mois – et donc présentant nécessairement moins de garanties et moins de services que la solution pérenne. Mais le débat n'est pas entre une solution pérenne et une solution transitoire, mais entre une solution transitoire et la situation actuelle, dont chacun connaît en particulier les délais d'accès et d'appariement avec les données de la base principale du SNDS. Un audit, mené par la DNS, est actuellement en cours s'agissant d'un projet spécifique porté par le HDH dans le cadre d'un appel à projets européen. La mission recommande d'élargir cet audit – et le cas échéant de le prolonger de quelques semaines – pour mener l'audit flash décrit ci-dessus.

A défaut de l'adoption rapide d'une solution transitoire, la « démultiplication » de copies, totales ou partielles, de la base principale du SNDS, qui est déjà à l'œuvre, risque de s'accélérer et, d'une part, présente le risque de « marginaliser » le HDH comme interlocuteur privilégié des demandes d'accès à la base principale du SNDS privées tant que celui-ci ne dispose pas d'une copie complète pour opérer des extractions et des appariements, et, d'autre part, constitue du point de vue de la mission une solution sous-optimale, qui n'est efficiente ni pour les acteurs de la recherche ni pour les pouvoirs publics. Mais cette « pente » est compréhensible : face au goulot d'étranglement des demandes d'accès à la base principale du SNDS au niveau de la CNAM, un certain nombre d'acteurs privés sont en effet tentés de s'adresser à des bureaux d'études qui entendent mettre à disposition des « blocs » du SNDS par pathologies, ou à des plateformes privées engagées dans des démarches d'homologation de systèmes fils SNDS. Il importe donc de définir sans tarder une solution transitoire permettant le transfert d'une copie de la base principale du SNDS, placée sous la responsabilité du HDH, afin de résorber la file d'attente des demandes d'accès traitées par la CNAM et afin d'accélérer les projets susceptibles de favoriser la recherche et les soins.

Recommandation n°2 Définir, dans les six mois, et sous pilotage du HDH, une solution transitoire pour accélérer notablement la mise à disposition des données, en particulier par la mise à disposition d'une copie de la base principale du SNDS.

Recommandation n°3 Procéder à un audit « flash » pour définir la solution transitoire souveraine la plus adaptée en l'état actuel, en incluant l'examen du centre d'accès sécurisé aux données (CASD) dans cet audit, afin de permettre le transfert d'une copie de la base principale du SNDS (cf. recommandation n°2).

2.2 Accélérer les délais de mise à disposition des données de santé et alléger les processus réglementaires

2.2.1 Alléger le processus réglementaire et généraliser les procédures simplifiées, pour que l'autorisation devienne l'exception

2.2.1.1 Préserver les procédures réglementaires simplifiées d'accès aux données de santé dans le cadre du futur règlement EHDS

Si la recherche à partir des données de santé devrait opérer à cadre réglementaire européen inchangé au moins jusqu'en 2027, la mission appelle l'attention des autorités françaises sur la nécessité de **préserver le modèle français des procédures réglementaires simplifiées d'accès aux données de santé dans le cadre du futur règlement EHDS**.

À cet égard, si ce dernier – à l'heure de la rédaction du présent rapport - ne comporte aucune disposition expresse prévoyant la possibilité pour les États membres d'instituer de telles procédures, il convient tout d'abord de relever que, dans son considérant 37, le projet de règlement EHDS précise son articulation avec le RGPD en matière d'utilisation secondaire des données de santé, en rappelant qu'« *il convient de se fonder sur les possibilités offertes par le règlement (UE) 2016/679 [RGPD] pour un droit de l'UE, dont les règles et les mécanismes prévoient*

des mesures appropriées et spécifiques pour garantir les droits et les libertés des personnes physiques ».

Ce même considérant prévoit que, lorsque le traitement de données est nécessaire à l'exécution d'une mission d'intérêt public, le responsable de traitement devra faire référence à une loi nationale ou un acte législatif du droit européen autres que le règlement EHDS qui lui confie le droit de traiter des données de santé à caractère personnel à des fins de conformité dans l'exercice de ses missions, ce qui laisse entendre que le recours des acteurs publics aux procédures simplifiées mises en œuvre sous l'empire du RGPD pourra perdurer.

En revanche, s'agissant des responsables de traitement privés, dont le traitement de données serait nécessaire à la poursuite d'intérêts légitimes, il est précisé que les garanties appropriées exigées par le RGPD devront être apportées par le règlement EHDS et qu'à ce titre, ce sont les décisions administratives de l'organisme responsable d'accès aux données qui détermineront leurs conditions d'accès aux données. Il y a donc lieu de s'interroger sur la possibilité, dans le cadre du futur règlement EHDS, au moins pour les acteurs privés, de continuer à bénéficier de la procédure de déclaration de conformité aux méthodologies de référence pour être autorisés à avoir accès à des données de santé à des fins d'utilisation secondaire, dès lors qu'une décision de l'organisme responsable d'accès aux données semble exigée par le règlement EHDS.

S'agissant par ailleurs des accès permanents, si le maintien dans sa forme actuelle du dispositif semble pour l'heure écartée par le règlement EHDS – l'article 48 du projet initial, relatif à la mise à disposition de données aux organismes du secteur public sans autorisation de traitement de données, ayant disparu des compromis successifs –, les autorités françaises indiquent avoir obtenu, à ce stade, l'insertion d'une disposition qui autoriserait les États membres à mettre en œuvre une **procédure accélérée d'autorisation pour les acteurs du secteur public**⁷⁶.

Toutefois, les finalités de cette procédure ne recouvrent pas celles liées à la recherche scientifique ayant trait aux secteurs de la santé ou des soins ou encore aux activités de développement et d'innovation pour les produits ou services contribuant à la santé publique ou à la sécurité sociale, ou à la garantie d'un niveau élevé de qualité et de sécurité des soins de santé, des médicaments ou des dispositifs médicaux.

La mission invite par conséquent les autorités françaises à continuer d'œuvrer auprès des institutions européennes pour que le maintien d'un dispositif d'accès permanent aux données du SNDS au bénéfice des acteurs publics, ou d'une solution s'en approchant le plus possible, demeure compatible avec le futur règlement EHDS.

À ce titre, l'examen du projet de règlement par le Parlement européen constituera une étape déterminante pour faire valoir l'intérêt de ces procédures pour l'accès des acteurs publics aux données de santé au titre de leurs missions d'intérêt public.

Il convient tout de même d'anticiper le cas de figure où le règlement européen remettrait en cause les procédures françaises d'accès permanent et de procédures simplifiées. À cet effet, une

⁷⁶ Cette procédure viserait les traitements répondant aux finalités mentionnées aux points a à c du paragraphe 1 de l'article 34 de ce règlement, c'est-à-dire : les activités pour des raisons d'intérêt public dans le domaine de la santé publique et de la santé au travail, notamment la garantie d'un niveau élevé de qualité et de sécurité des soins de santé et des médicaments ou dispositifs médicaux ; aider les organismes du secteur public ou les institutions, organes et organismes de l'Union, dont les autorités réglementaires, dans le secteur de la santé ou des soins, à accomplir les tâches inscrites dans leur mandat ; et produire des statistiques officielles en rapport avec les secteurs de la santé ou des soins.

option, évoquée lors des ateliers organisés par la mission, serait de **mobiliser le mécanisme des autorisations uniques au bénéfice des acteurs publics de la recherche** afin de leur permettre de continuer à effectuer les traitements auxquels ils procèdent dans le cadre de leurs accès permanents.

Cette réflexion mériterait d'être poursuivie et élargie, pour l'ensemble des voies d'accès simplifiées qui pourraient être remises en cause.

Recommandation n°4 Mobiliser les leviers de négociation disponibles pour ajuster le projet de règlement EHDS, notamment lors de son examen par le Parlement européen, afin de préserver les mécanismes d'accès permanent et de procédures simplifiées ; anticiper l'éventualité d'une mise en cause potentielle de ces mécanismes en élaborant de nouvelles modalités d'accès simplifié, avec des effets équivalents.

2.2.1.2 Rénover le processus d'élaboration des méthodologies de référence pour mieux répondre aux besoins des utilisateurs

Par ailleurs, pour mieux répondre aux besoins de l'écosystème, l'élaboration et l'actualisation des référentiels simplifiés et des méthodologies de référence doivent s'appuyer sur un processus de recueil plus dynamique des souhaits des utilisateurs de données. La mission salue à cet égard la concertation que compte engager la CNIL fin 2023-début 2024, en collaboration avec le HDH, pour recueillir directement les attentes en termes d'évolution des premières méthodologies de référence et du référentiel des entrepôts des données de santé. La mission estime néanmoins que le processus d'élaboration des référentiels simplifiés et des méthodologies de référence appelle une refonte plus substantielle.

Une première option pourrait consister à maintenir le circuit actuel d'élaboration des référentiels simplifiés, à la main de la CNIL moyennant une concertation avec le HDH, les utilisateurs et les membres du Comité stratégique des données de santé.

Une seconde option, qui a la préférence de la mission, serait de **confier au Comité stratégique des données de santé un rôle d'impulsion des évolutions réglementaires des procédures simplifiées**. Il pourrait, à ce titre, s'appuyer sur le HDH pour opérer une remontée des attentes des acteurs en matière d'évolutions réglementaires. En sa qualité de GIP regroupant les représentants de l'ensemble des parties prenantes, publiques et privées, du secteur des données de santé, le HDH serait ainsi chargé de centraliser le recueil des demandes ponctuelles d'évolutions réglementaires mais aussi de sonder périodiquement la communauté sur les besoins d'évolution. Fort des retours d'expérience qui lui seraient transmis par le HDH, le Comité stratégique aurait alors la mission de proposer de nouveaux référentiels simplifiés et méthodologies de référence ou des projets d'actualisation de ces textes, préparés au sein de groupes de travail, au ministre chargé de la santé, lequel les validerait – publication par arrêté – après avis de la CNIL.

Cette évolution du processus de l'élaboration des référentiels simplifiés et méthodologies de référence nécessiterait une évolution législative puisque le II de l'article 66 de la loi « informatique et libertés » confie aujourd'hui la responsabilité de cette élaboration à la CNIL.

La méthode consistant à confier à une instance administrative de concertation le soin de proposer des référentiels ou lignes directrices opposables, sous réserve de leur approbation par le pouvoir

réglementaire, n'est pas inhabituelle, comme le montrent les exemples du comité national de prévention et de santé au travail et du comité national pour l'emploi⁷⁷.

En anticipation de cette évolution législative, la CNIL pourrait dès à présent déléguer au comité stratégique, avec l'appui du Health data hub, le soin de coordonner les travaux d'élaboration des méthodologies de référence, tout en conservant la validation formelle de ces textes.

Recommandation n°5 Changer le mode d'élaboration des méthodologies de référence, en prévoyant leur adoption par arrêté du ministre chargé de la santé après avis de la CNIL, en confiant au HDH la mission de recueillir les attentes de l'écosystème en matière d'évolution de ces méthodologies et au Comité stratégique des données de santé la mission de proposer des projets de référentiels simplifiés et de méthodologies de référence.

2.2.1.3 Simplifier l'architecture des méthodologies de référence

La mission s'interroge, par ailleurs, sur le nombre pertinent de méthodologies de référence – déjà au nombre de cinq concernant les recherches n'impliquant pas la personne humaine –. Plutôt que la multiplication de méthodologies pour répondre à un type déterminé de recherches correspondant à des finalités bien précises et destinées à des catégories restreintes de responsables de traitement, la mission recommande la **définition de méthodologies de référence « cadres » plus lisibles pour les acteurs de la recherche**.

À titre d'exemple, la question se posera, à terme, de la nécessité de maintenir une méthodologie de référence réservée aux établissements de santé publics et privés et aux fédérations hospitalières pour l'accès aux données du PMSI, alors que de nouvelles méthodologies de référence (MR-007 et MR-008) sont intervenues pour encadrer l'accès des acteurs publics et privés de la recherche aux données de la base principale du SNDS, respectivement au titre de leurs missions d'intérêt public et de leurs intérêts légitimes. Il y aurait matière à regrouper ces méthodologies de référence, en harmonisant notamment certaines conditions déterminantes de l'accès aux données (consultation obligatoire ou pas du CESREES, possibilité de chaînage de données issues du SNDS avec d'autres données...).

Recommandation n°6 Rationaliser le paysage des méthodologies de référence en privilégiant le format des méthodologies de référence "cadres".

⁷⁷ En matière de santé et de sécurité au travail, le comité national de prévention et de santé au travail est ainsi chargé de proposer les référentiels et les principes guidant l'élaboration du cahier des charges de certification des services de prévention et de santé au travail interentreprises, ces référentiels et principes devant ensuite être validés par voie réglementaire.

En matière de structuration du service public de l'emploi, le comité national pour l'emploi, institué par le projet de loi relatif au plein emploi adopté définitivement par l'Assemblée nationale le 14 novembre 2023, est chargé de définir un socle commun de services au bénéfice des personnes et des employeurs, et d'établir des méthodologies et des référentiels comportant, le cas échéant, des objectifs de qualité de service ainsi qu'un cahier des charges recensant les besoins des membres du réseau pour assurer l'interopérabilité de leurs systèmes d'information, étant entendu que les actes adoptés à ce titre doivent être approuvés par le ministre chargé du travail avant leur publication.

2.2.1.4 Poursuivre les assouplissements réglementaires notamment sur l'utilisation du NIR et les entrepôts de données de santé

En outre, un des principaux manques identifiés par les acteurs interrogés par la mission dans les actuelles méthodologies de référence est **l'impossibilité de recueillir et de traiter, dans le cadre de ces méthodologies de référence, le numéro d'inscription au répertoire national d'identification des personnes physiques (NIR)**, ce qui limite fortement les possibilités pour les projets de recherche recourant à une méthodologie de référence d'apparier des données issues de différentes bases de données.

Si les catégories de responsables de traitement et les finalités des traitements impliquant des données comportant le NIR sont déterminées par un décret en Conseil d'État⁷⁸, l'article 30 de la loi « Informatique et libertés » prévoit que ne sont pas soumis à cette contrainte les traitements portant sur des données parmi lesquelles figure le NIR ou qui requièrent une consultation de ce répertoire qui ont exclusivement des **finalités de recherche scientifique ou historique**, sous réserve que le NIR fasse, dans ce cas, l'objet d'une pseudonymisation (« *opération cryptographique lui substituant un code statistique non signifiant* »).

Jusqu'ici, les traitements requérant une utilisation du NIR font l'objet d'une autorisation préalable de la CNIL, laquelle a diffusé plusieurs **fiches pratiques à destination des responsables de traitement sur les conditions d'utilisation du NIR**, aux fins d'acculturer les acteurs de la recherche à l'utilisation d'une donnée aussi sensible.

Signe d'une ouverture de la CNIL à une utilisation plus souple du NIR dans le cadre de bases de données, la possibilité de collecter l'identifiant national de santé (INS) – qui, aux termes de l'article R. 1111-8-1 du code de la santé publique, correspond au NIR – est désormais admise pour les entrepôts de données de santé dans le cadre du référentiel adopté par la CNIL en octobre 2021. Cette collecte n'est en revanche pas autorisée dans le cadre des essais cliniques et des cohortes.

Les méthodologies de référence gagneraient à évoluer pour permettre le chaînage de données entre plusieurs bases, dont celles du SNDS, par un **recours facilité au NIR** dès lors que les exigences dans la manipulation d'une donnée aussi sensible que le NIR, telles qu'elles figurent dans les fiches pratiques de la CNIL, sont intégrées dans les méthodologies de référence.

Dans ces conditions, la mission recommande d'étendre le champ des méthodologies de référence pour mieux répondre aux besoins des acteurs, en **facilitant notamment, dans le cadre de ces méthodologies, la réalisation de projets de recherche nécessitant des appariements**, sous réserve du respect de conditions exigeantes dans la manipulation du NIR.

Recommandation n°7 Étendre le champ des méthodologies de référence pour notamment permettre le recueil et le traitement du NIR et le chaînage de données issues de plusieurs bases dans le cadre de projets de recherche.

Le référentiel sur les entrepôts de données de santé nécessite également des ajustements. En premier lieu, **la définition d'un entrepôt de données de santé nécessite d'être clarifiée**, pour

⁷⁸ Décret n° 2019-341 du 19 avril 2019 relatif à la mise en œuvre de traitements comportant l'usage du numéro d'inscription au répertoire national d'identification des personnes physiques ou nécessitant la consultation de ce répertoire.

distinguer clairement les entrepôts de données de santé des cohortes ou les registres, qui relèvent d'approches le plus souvent très différentes des EDS. Le cadre réglementaire des cohortes et des registres mériterait par ailleurs aussi d'être mieux défini et précisé au regard des exigences du RGPD et de la CNIL, avec une approche adaptée. À ce titre, les discussions actuellement menées entre la CNIL et l'INSERM pour mieux cerner et définir les différentes cohortes selon leurs typologies sont intéressantes. La mission recommande de prolonger cette démarche jusqu'à l'élaboration d'un référentiel pour les cohortes.

Pour les EDS, la mission préconise de s'appuyer sur la définition de la Haute Autorité de santé pour laquelle « *les entrepôts de données de santé (EDS) désignent la mise en commun des données d'un ou plusieurs systèmes d'information médicaux, sous un format homogène pour des réutilisations à des fins de pilotage, de recherche ou dans le cadre des soins* »⁷⁹.

Par ailleurs, de nombreux acteurs ont souligné que les exigences définies par le référentiel EDS étaient particulièrement ambitieuses, d'autant qu'elles se cumulent avec les exigences du référentiel Hébergement de santé (HDS), le cas échéant avec celles du référentiel SNDS et du référentiel SecNumCloud. Le haut degré d'exigence est lié au fait que le référentiel vise l'ensemble des entrepôts, y compris les plus gros et les plus sensibles. Il paraît d'ailleurs surprenant que les grands entrepôts, les plus sensibles en termes de contenu de données et de sécurité, ne fassent plus l'objet d'autorisations au moment de leur création, même si en pratique la plupart sollicitent une telle autorisation. Une refonte du référentiel pourrait donc être envisagée, en visant un allègement du niveau d'exigence, tandis qu'une procédure d'autorisation serait maintenue *a minima* pour les plus gros entrepôts pour lesquels un niveau d'exigence plus élevé pourrait se justifier.

De manière générale, la mission préconise de généraliser cette logique de réserver les autorisations aux cas les plus sensibles.

Enfin, un travail spécifique devrait être initié sur les **modalités d'application des principes d'anonymisation et de minimisation** pour les projets de recherche scientifique s'appuyant sur les données de santé. Un assouplissement des dispositions de droit français transposant l'article 89 du RGPD, qui prévoit des possibilités de dérogation au principe d'anonymisation et de minimisation, mériterait d'être envisagé. A défaut, une réécriture plus ouverte de l'article 89 du RGPD pourrait être mise à l'ordre du jour par la France.

2.2.1.5 Étendre le champ des méthodologies de référence pour que l'autorisation devienne l'exception

L'ensemble des recommandations relatives au processus réglementaire doit permettre de **consacrer le recours aux référentiels simplifiés et aux méthodologies de référence comme modalité d'accès de droit commun aux données de santé à des fins de recherche**, et conduire concrètement à une diminution significative du nombre de demandes d'autorisation, celles-ci devant se cantonner d'une part aux projets de création d'un entrepôt de données de santé, d'autre part aux projets de recherche les plus complexes et les plus sensibles. Il s'agit donc d'une inversion de la logique qui prévalait jusqu'alors, l'autorisation devenant l'exception, et la

⁷⁹ Haute Autorité de santé, *Entrepôts de données de santé hospitaliers en France – Quel potentiel pour la Haute Autorité de santé ?*, rapport validé par le collège le 20 octobre 2022.

procédure simplifiée la règle. Le recours aux procédures d'autorisations uniques permettrait aussi de contribuer à la diminution de demandes d'autorisation.

En 2022, 286 projets de traitements de données de santé à des fins de recherche, d'étude ou d'évaluation n'impliquant pas la personne humaine ont encore fait l'objet d'une autorisation de la CNIL. Dans une logique de renforcement du recours aux procédures simplifiées, l'objectif – indicatif – que la mission propose de fixer est que dans un délai de trois ans, ce nombre d'autorisations soit inférieur à 100 par an.

Recommandation n°8 Fixer un objectif quantifié de diminution du nombre d'autorisations délivrées par la CNIL pour que, de manière effective, l'autorisation devienne l'exception par rapport aux procédures simplifiées.

2.2.1.6 Assouplir l'obligation de ré-information individuelle des patients en contrepartie d'une information renforcée

L'obligation d'une ré-information individuelle du patient à l'occasion de chaque réutilisation de ses données de santé dans le cadre d'un projet de recherche mobilisant la base dans laquelle ses données ont été collectées (EDS, cohorte, registre...) constitue souvent une charge regardée comme excessive par le porteur de projet, qui sollicite alors une exception à cette obligation sur le fondement de l'article 14 du RGPD (paragraphe 5, point *b*) dans le cadre d'une étude rétrospective, en invoquant l'impossibilité de réaliser cette information ou les efforts disproportionnés qu'elle impliquerait. La démonstration de cette impossibilité ou du caractère disproportionné de ces efforts reste néanmoins, en elle-même, une contrainte pour le porteur de projet.

La mobilisation *quasi*-systématique de cette exception révèle le **caractère inadapté de l'obligation de ré-information individuelle du patient**, inscrite à l'article 69 de la loi « Informatique et libertés », dans le cas des réutilisations des données de santé dans le cadre de nouveaux projets de recherche, *a fortiori* concernant des bases de données anciennes, pour lesquelles la récupération du consentement des années après leur constitution est impossible. À cet égard, il semble nécessaire d'assouplir ou, du moins, de clarifier la portée des dispositions des articles 13 (paragraphe 3) et 14 (paragraphe 5, point *b*) du RGPD relatives aux conditions d'information des patients en cas de réutilisation de leurs données de santé pour de nouvelles finalités, à partir du moment où cette réutilisation continue de s'inscrire dans une démarche de recherche, d'étude ou d'évaluation en santé.

À cet effet, le projet de règlement EHDS pourrait être l'occasion de prévoir que, par dérogation à ces dispositions du RGPD, **une réutilisation par un acteur public des données de santé à caractère personnel à des fins de recherche est par principe possible, sans nécessité d'une ré-information individuelle par le responsable de traitement**, dès lors que le patient a été informé, au moment de la collecte de ses données de santé, de la possibilité qu'elles soient réutilisées à des fins de recherche par des acteurs publics et qu'il ne s'est pas opposé à cette réutilisation, sous réserve que le responsable de traitement mette à disposition du public (portail de transparence, affiches dans les établissements de santé...) les informations requises par le RGPD.

Dans sa version connue de la mission, le compromis sur le projet de règlement EHDS prévoit :

- une obligation d'information individuelle des personnes concernées par le détenteur des données de santé, mais aménage une exception similaire à celle prévue à l'article 14 du RGPD en cas d'impossibilité de fournir à chaque personne concernée les informations requises ou si cette information individuelle impliquerait un effort disproportionné ;
- une obligation d'information des personnes concernées par les organismes d'accès aux données de santé (HDAB). S'il est prévu que les HDAB doivent rendre publiques et facilement consultables par des moyens électroniques les conditions dans lesquelles les données de santé électroniques sont mises à disposition pour un usage secondaire⁸⁰, il est néanmoins laissé la possibilité aux États membres de prévoir dans leur droit national l'obligation de fournir des informations individualisées à chaque personne concernée au titre de l'article 14 du RGPD.

Mais il est difficile, à ce stade, de déterminer les conditions dans lesquelles les obligations d'information pesant sur le détenteur des données de santé ou le HDAB s'articulent avec celles qui pèsent sur le responsable du traitement qui réutilise les données mises à disposition en application de l'article 14 du RGPD. Dès lors, il serait utile de clarifier que la mise à disposition auprès du public par le détenteur des données ou par le HDAB, par exemple par la voie d'un portail de transparence, des informations concernant la réutilisation des données par des acteurs publics suffit à satisfaire l'obligation d'information des personnes concernées pesant sur ces responsables de traitement publics.

Une telle clarification du droit communautaire, spécifique pour la réutilisation des données de santé à des fins de recherche, permettrait de **formaliser en droit interne le principe selon lequel, sauf refus explicite manifesté dès la collecte de ses données, le patient est présumé avoir consenti à la réutilisation de celles-ci par des acteurs publics à des fins de recherche en santé**, selon une logique analogue à celle qui prévaut en matière de consentement au don d'organes.

Recommandation n°9 Clarifier, si besoin dans le projet de règlement EHDS que, par dérogation à l'obligation du RGPD de ré-information individuelle du patient à chaque réutilisation de ses données de santé, le patient, qui a déjà été informé, à l'occasion de la collecte de ses données, de la possibilité que ses données soient réutilisées, n'a pas à être à nouveau informé.

La contrepartie de l'absence de ré-information individuelle par des acteurs publics à chaque réutilisation serait une transparence renforcée – par exemple, *via* le portail de transparence centralisé par le HDH sous réserve de sa faisabilité (voir partie 2.4.4.1) –, sur chaque projet de recherche ayant eu accès aux données des personnes concernées mais également une **simplification des conditions d'exercice par ces dernières de leurs droits d'« opt-out » (voir partie 2.4.4.2)**, leur permettant de s'opposer à la réutilisation de leurs données selon le projet de recherche.

⁸⁰ C'est-à-dire les autorisations d'accès délivrées et leur descriptif.

2.2.1.7 Renforcer l'accompagnement et la formation des acteurs à la réglementation et aux procédures simplifiées

Dans une logique de **renforcement de la confiance dans les porteurs de projet de recherche**, la mission recommande de favoriser, conformément à la logique inscrite dans le RGPD, la **responsabilisation des responsables de traitement**, dans le respect des règles applicables à l'utilisation des données de santé à caractère personnel, par un recours renforcé aux méthodologies de référence. La déclaration de conformité à ces méthodologies doit permettre au responsable de traitement, conformément à l'article 5, point 2 du RGPD, de démontrer que les exigences de ce dernier sont bien respectées.

Cela implique de **parfaire l'appropriation par les acteurs de la recherche des procédures simplifiées existantes**, un certain nombre de demandes d'autorisation dont est saisie la CNIL pouvant en réalité relever d'une méthodologie de référence ou être satisfaites en exploitant le dispositif de l'accès permanent ou le mécanisme de la décision unique d'autorisation. À cet égard, deux leviers doivent en particulier être actionnés :

- il est impératif de renforcer la formation tant des porteurs de projet que des équipes juridiques et des DPO des institutions dont ils relèvent à l'environnement réglementaire de la recherche en santé, afin que les chercheurs soient aiguillés le plus tôt possible vers la procédure réglementaire la plus adaptée à leur projet ;
- reconnu pour la qualité de son accompagnement, le HDH pourrait être mobilisé pour améliorer la pédagogie sur les outils réglementaires disponibles, en collaboration avec la CNIL, pour lever un certain nombre d'incompréhensions sur le champ des possibles permis par ces outils, et conseiller les porteurs de projet et leurs équipes juridiques sur la procédure réglementaire la plus pertinente, ou plus généralement sur la mise en conformité avec le RGPD. Le HDH pourra s'appuyer sur les différents guides pratiques réalisés par la CNIL (comme le guide pratique sur les modalités de circulation du NIR), à la réalisation desquels il pourrait désormais être associé.

Recommandation n°10 Renforcer la formation des porteurs de projet, des équipes juridiques et des DPO à la réglementation applicable en matière de recherche en santé et mobiliser le HDH pour les conseiller sur la procédure réglementaire la plus adaptée au projet.

2.2.2 Simplifier le processus d'évaluation scientifique et éthique, dans une logique de subsidiarité

Compte-tenu de la mission confiée en juillet 2023 au comité consultatif national d'éthique pour les sciences de la vie et de la santé, portant sur l'organisation de l'évaluation éthique pour les recherches impliquant la personne humaine, la mission n'a pas traité des procédures éthiques et scientifiques sur ce champ. Elle s'est ainsi concentrée sur les demandes d'autorisation ou d'études n'impliquant pas la personne humaine.

Pour ces dernières, l'article 76 de la loi « informatique et libertés » impose aujourd'hui un avis préalable du CESREES, portant sur la méthodologie et la pertinence scientifique et éthique d'un projet ainsi que sur sa finalité d'intérêt public, avant qu'une autorisation de traitement soit accordée par la CNIL, permettant l'accès à des données de santé.

On a vu précédemment (cf. a) de la partie 1.2.3.1) que les projets de recherches pouvaient faire l'objet d'examens par plusieurs comités scientifiques ou éthique, ce qui suscite des incompréhensions légitimes et des délais supplémentaires.

L'approche française tranche avec celle privilégiée dans d'autres pays où l'intervention d'un comité éthique est généralement réservée aux projets de recherche qui, après **procédure de tri** par l'autorité responsable de la base requêtée, sont identifiés comme présentant des spécificités justifiant une évaluation scientifique et éthique dédiée. Cette procédure de tri intervient par exemple s'agissant de l'accès aux données de la base britannique CPRD ou des demandes traitées par le FINDATA en Finlande qui ne sollicite l'avis d'un comité scientifique et éthique qu'en cas de besoin. De même, l'expertise du centre Ethox sur les demandes d'accès aux données de la UK Biobank n'est mobilisée qu'en cas de besoin.

Dès lors, et sans préjudice du fonctionnement actuel globalement satisfaisant du CESREES, la mission a examiné plusieurs options pour alléger le processus actuel, dans une logique de renforcement de la confiance faite aux acteurs et en particulier aux chercheurs.

Une première option consisterait en une suppression de l'examen des projets par le CESREES. Cela consacrerait la responsabilisation du responsable de traitement auquel il appartiendrait, comme c'est du reste déjà le cas dans le cadre des traitements déclarés conformes aux MR-004, MR-005 et MR-006⁸¹, de justifier de la validation scientifique de son protocole de recherche et de l'intérêt public du projet en cas de contrôle *a posteriori* par la CNIL.

La suppression pure et simple de l'examen du CESREES pose néanmoins une difficulté à l'égard des traitements mis en œuvre par des acteurs privés dont la vérification de la finalité d'intérêt public peut appeler une vigilance particulière. C'est pourquoi la mission a écarté cette option.

Une autre option pourrait être de **dispenser les recherches portées par des chercheurs publics d'un examen préalable par le CESREES, dès lors que leurs recherches font l'objet d'un avis par un comité scientifique et éthique local**, dont se sont dotées les principales structures de recherche universitaire et hospitalière. **Cette dispense serait conditionnée au respect par le CSE d'un cahier des charges défini nationalement, fixant les principes de leur gouvernance et de leur fonctionnement, afin de garantir l'homogénéité des procédures.**

À cet égard, il convient de relever que le Comité stratégique des données de santé a mis en place en décembre 2022 un groupe de travail, consacré à la gouvernance de l'accès aux données de santé, chargé de réfléchir à une **harmonisation du traitement des demandes d'accès aux données et de l'organisation des CSE au sein des bases de données locales (entrepôts, registres, cohortes)**. À ce stade, il a livré un « kit » de démarrage pour les CSE des EDS, qui intègre notamment une trame de règlement intérieur, une fiche d'évaluation et des modèles d'avis. Les critères

⁸¹ En revanche, l'avis du CESREES a été maintenu pour les MR-007 et MR-008, la CNIL mettant en avant la nécessité de s'assurer, dans un premier temps, de la maturité des acteurs en contrepartie de la profondeur historique importante et des finalités très larges consenties dans l'accès aux données du SNDS dans ces deux méthodologies de référence.

d'évaluation sont ainsi précisés et les délais de délivrance des avis des CSE sont également encadrés (accord tacite dans un délai d'un mois, renouvelable une fois).

Cette production, d'abord focalisée sur les EDS, constitue une première étape précieuse dans l'harmonisation de la gouvernance et du fonctionnement des CSE locaux, qui mériterait d'être étendue, dans le cadre d'un cahier des charges national, à l'ensemble des CSE créés auprès de bases de données locales ou nationales (EDS, cohortes, registres).

L'élaboration du cahier des charges serait confiée au Comité stratégique des données de santé, le groupe de travail qu'il a mis en place sur la gouvernance de l'accès aux données de santé comprenant d'ores et déjà des représentants des principales parties prenantes des principaux types de bases de données locales ou nationales requêtées⁸². Le CESREES gagnerait à être associé à l'élaboration de ce cahier des charges.

Le cas échéant, le CSE local mobilisé pourra être celui mis en place à un niveau inter-hospitalier ou régional, comme c'est le cas de l'entrepôt Ouest Data Hub porté par le GCS HUGO.

Cette recommandation devra s'accompagner d'une montée en compétences des CSE locaux sur la connaissance du SNDS, pour garantir la qualité des avis et l'homogénéité des processus d'évaluation entre les établissements pour les projets qui mobilisent des données du SNDS. L'effort de formation (voir infra, partie 2.2.3.2) devra donc plus particulièrement concerner les membres des CSE locaux.

Par ailleurs le respect des délais des CSE inscrit dans le cahier des charges national devra être suivi, pour que la dispense d'avis du CESREES se concrétise réellement par une accélération des délais d'examen des projets sur un plan scientifique et éthique pour les chercheurs.

Recommandation n°11 Dispenser les projets de recherche n'impliquant pas la personne humaine d'un avis du CESREES, dès lors qu'un comité scientifique et éthique local respectant un cahier des charges national a rendu un avis.

S'agissant des projets de recherche impliquant l'appariement de données du SNDS avec celles issues d'autres bases de données, locales ou nationales, la mission recommande d'étendre aux MR-007 et MR-008 la logique de dispense d'avis du CESREES mais aussi la possibilité de chaînage des données du SNDS avec d'autres données. En particulier, les titulaires d'un accès permanent au SNDS devraient être dispensés d'obtenir un avis du CESREES en cas de recherche nécessitant l'appariement de données du SNDS avec celles d'autres bases de données pour lesquelles ils ont déjà obtenu un avis favorable d'un CSE local.

⁸² CHU, DGRI, registre « REIN » de l'Agence de biomédecine, CESREES, France Cohorte, Cohorte ESME d'Unicancer...

2.2.3 Accélérer la mise à disposition des données de la base principale du SNDS

2.2.3.1 La priorisation de l'accès pour certains projets de recherche d'intérêt majeur, complexe à mettre en œuvre, n'est pas souhaitée par les acteurs

L'option d'une priorisation de l'ensemble des demandes d'accès aux données du SNDS a été écartée par la mission, la hiérarchisation de l'ensemble de ces demandes et la définition de critères légitimes et pertinents apparaissant rapidement trop complexe.

La mission a également étudié l'option de l'attribution par le CESREES d'un label aux projets de recherche répondant à une priorité particulièrement stratégique dans le domaine de la recherche en santé, **leur permettant d'être identifié comme « urgent »** et leur octroyant la garantie d'un accès aux données du SNDS dans un délai accéléré.

Ce label ne serait attribué que si le projet s'inscrit dans des **priorités stratégiques d'intérêt majeur pour la recherche en santé**, qui pourraient être définies conjointement par l'Agence de l'innovation en santé (AIS), le HDH et le Comité stratégique des données de santé, en concertation avec les organismes de recherche nationaux.

Il ressort néanmoins des auditions menées par la mission que la mise en place d'un tel processus de « *fast track* » pour l'accès aux données du SNDS suscite de fortes réticences. D'une part, l'identification d'aires thérapeutiques prioritaires qui justifieraient un accès facilité au SNDS et la définition de critères, admis par l'ensemble de la communauté de recherche, pour l'identification des priorités stratégiques constitueraient un exercice hautement périlleux. Par ailleurs, les risques d'iniquité dans le traitement des demandes d'accès sont avancés, avec une crainte que les projets de nature privée ne soient désavantagés. De manière générale, la crainte que, de manière indirecte, la création d'un processus accéléré ne conduise à entraver définitivement l'accès aux données pour une partie des autres projets de recherche jugés non prioritaires est formulée par de nombreux acteurs. La mission prend acte de ces craintes, tout en constatant que d'une manière ou d'une autre une forme de priorisation mérité d'être définie, et recommande que le Comité stratégique des données de santé puisse se saisir de ce sujet en concertation avec les acteurs.

2.2.3.2 La mise à disposition des données de la base principale du SNDS doit être repensée dans une logique de responsabilisation des acteurs

On a vu précédemment que le temps d'accès élevé aux données du SNDS s'explique notamment par le délai élevé de mise à disposition des données par la CNAM, avec un délai moyen de 10 à 12 mois, en augmentation (cf partie 1.2.3.1).

Deux options ont été examinées pour réduire de manière significative ces délais à courte échéance (sans qu'il soit nécessaire d'attendre la mise à disposition de la copie du SNDS auprès du Health data hub). **L'objectif recherché est la diminution forte du temps de ciblage de la population entre le porteur de projet et la CNAM**, qui peut nécessiter de nombreux aller-retours avec les porteurs de projet pour préciser le périmètre exact des données, au-delà de l'expression des besoins validée par le CESREES et la CNIL. Ces aller-retours sont justifiés par la CNAM par le souci de transmettre des données exploitables par les demandeurs, qui pourront

répondre aux besoins de leur étude, mais aussi par le souci du respect de principe de minimisation des données. Pour les porteurs de projet, ces aller-retours sont perçus comme longs et redondants avec les avis formulés par la CNIL et le CESREES sur le périmètre des données sollicité dans l'expression de besoin. Par ailleurs, ce mode de fonctionnement entre en contradiction avec la logique du RGPD de responsabilisation des responsables de traitements de données.

La première option, proposée par le Health data hub, est la mise à disposition de « blocs » du SNDS pour répondre aux demandes d'accès aux données sans appariements. La logique serait de constituer des copies de la base principale du SNDS sur des périmètres de populations spécifiques couramment utilisés dans les projets. Une quarantaine de blocs serait ainsi constituée, sur des périmètres comme le diabète ou les maladies cardiovasculaires, etc. L'intégralité du bloc serait alors mis à disposition du porteur de projet, à qui il reviendrait de sélectionner sa population d'intérêt au regard des critères soumis aux CESREES et à la CNIL et de s'assurer que cette dernière répond à son besoin. Le principe de minimisation serait donc déporté sur le porteur de projet.

Cette option a toutefois été écartée par la mission, pour les raisons suivantes :

- de nombreuses demandes seront difficiles à satisfaire par les blocs : si aucun bloc ne permet de répondre à la demande, ou, s'il est nécessaire d'accéder à plusieurs blocs, leur appariement sera très complexe (risque de doublons,...) ;
- un risque d'élargissement progressif de la volumétrie des blocs (pour répondre au mieux aux demandes des acteurs), jusqu'à constituer des copies pérennes de sous-parties importantes du SNDS.

La seconde option, retenue par la mission, consisterait à mettre à disposition, de manière temporaire, un accès ponctuel aux données du SNDS pour le porteur du projet, pour que celui-ci réalise le découpage des variables et de la population (à la place des équipes de la CNAM). Cela conduit, comme dans l'option précédente, à déléguer la responsabilité de la minimisation au porteur de projet mais de manière plus substantielle, et sans recréer de copies pérennes de la base. La logique serait similaire à la création temporaire d'un accès permanent pour les porteurs de projets.

Cette option concernerait en premier lieu les demandes d'extraction de données sans appariements (qui représentent 75% des traitements réalisés par la CNAM entre 2019 et 2023. Sa mise en œuvre est souhaitable également pour les demandes d'appariements indirects, ce qui nécessite la mise à disposition de bulles sécurisées pour charger la base à apparier. Pour ceux deux configurations (extraction sans appariement et appariement indirect), cela suppose que les porteurs de projets aient une bonne connaissance du SNDS, ou qu'ils soient accompagnés par des acteurs experts. L'étape d'expression de besoins serait maintenue pour la demande auprès de la CNIL et du CESREES (ou d'un CSE local).

Cette option, qui recueille un consensus favorable des principaux acteurs concernés, pourrait être testée à brève échéance auprès d'un premier réseau d'acteurs habitués à manipuler les données du SNDS, et dont la compétence est reconnue, sous réserve du respect d'une charte rappelant la responsabilité de minimisation et les engagements des acteurs. Ce test permettra de préciser quelles opérations précises peuvent être déléguées et de déterminer les conditions informatiques adéquates.

La mise en œuvre de cette option doit s'accompagner d'une meilleure transparence sur les délais de mise à disposition des données du SNDS, afin de mieux informer les porteurs de projets sur les délais d'accès qu'ils peuvent escompter en fonction de la nature de leur projet, et de fournir une information sur la taille de la file d'attente de projets. Ceci permettra également d'assurer un suivi sur l'évolution des délais de mise à disposition des données à travers le développement d'indicateurs de délais. La définition des bons indicateurs de délai mérite une instruction, même rapide.

A terme, de tels indicateurs pourraient être développés et prévus pour l'accès à l'ensemble des bases de données, et pas seulement à la base principale du SNDS. En effet, un suivi des délais d'accès aux données sera rendu nécessaire par le futur règlement EHDS, qui prévoit des délais contraints pour la mise à disposition des données. Le comité stratégique des données de santé pourrait ainsi organiser une concertation pour définir un (ou des) indicateurs communs aux bases de données pour mesurer les délais d'accès aux données, en anticipation du règlement européen.

Recommandation n°12 Avancer collectivement et rapidement sur la délégation au porteur de projet de la responsabilité de la minimisation, avec une mise à disposition temporaire d'un accès à l'ensemble de la base principale du SNDS permettant au porteur de procéder aux opérations de préparation des données nécessaires à l'étude, afin d'accélérer très significativement l'accès aux données du SNDS.

Recommandation n°13 Développer à brève échéance un indicateur de délai sur la mise à disposition des données du SNDS ; à moyen terme, étendre cette logique de transparence à l'ensemble des bases de données en santé avec l'élaboration concertée d'indicateurs de délai au sein du comité stratégique des données de santé.

2.2.3.3 Favoriser l'exploitation des données de la base principale du SNDS à travers le renforcement de formations, une standardisation des données et le développement d'outils de « bac à sable »

Tout d'abord, l'effort de formation à la base principale du SNDS doit être accentué et davantage décentralisé, dans une logique de démultiplication, en s'appuyant sur des relais permettant de réaliser des formations auprès des chercheurs ou porteurs de projet en proximité. Ces relais, qui conduiront à ce que l'ensemble des formations ne soient plus assurés uniquement par la CNAM, pourront être assurés par le Health data hub (qui a d'ores et déjà déployé des efforts importants pour renforcer les outils de formation à la base principale du SNDS), ou encore par de grands organismes de recherche comme l'INSERM ou par les entrepôts de données de santé hospitaliers. Il s'agit tant de développer la formation initiale, avec le développement de modules de formation au SNDS dans le milieu universitaire, que de développer des formations continues, en particulier pour les membres des CSE locaux, dont la connaissance du SNDS doit être renforcée pour pouvoir se prononcer sur les projets d'appariements avec le SNDS (cf *supra*). Les membres des comités de protection des personnes (CPP) pourraient également utilement être formés sur ce sujet.

Les formations devront notamment permettre de renforcer la qualité des expressions de besoins de données de la base principale du SNDS et d'accompagner les porteurs de projet pour la minimisation des données. Des formations en ligne gagneraient à être développées, à l'instar des modules de formation développés par le Health data hub.

La stratégie d'accélération du numérique en santé comporte un axe important sur la formation, avec un investissement total de 19 M€. La mission recommande qu'une partie de ces crédits soit consacré au développement de formation à l'utilisation de la base principale du SNDS.

Recommandation n°14 Développer les formations à l'utilisation de la base principale du SNDS, dans une logique de démultiplication et de « décentralisation », avec un financement par la stratégie d'accélération du numérique en santé.

Une partie du besoin de formation est liée à la complexité de la structuration de la base principale du SNDS. Des actions sont à mener pour convertir la base principale dans un standard facilement requêtable avec à terme la possibilité d'interfaces de programmation d'application pour requêter cette base. Ces travaux ont déjà été initiés par le Health Data Hub à travers la conversion de la base principale au format OMOP.

Dans le prolongement d'une réflexion qui a émergé des acteurs lors des ateliers participatifs organisés par la mission, il est également préconisé de **développer différents outils pour favoriser l'exploitation des données, dans un premier temps de la base principale du SNDS puis dans un second temps de tous les producteurs de données de santé qui le souhaiteraient, par des chercheurs ou des porteurs de projets qui souhaitent « tester » les données et confirmer la faisabilité de leur projet.** Au-delà de la mobilisation plus importante de l'échantillon du SNDS ESND, accessible rapidement sans avis de la CNIL ou du CESREES (référentiel de la CNIL), il a été suggéré le développement ou la consolidation de trois types d'outils dans une logique de « bac à sable » :

- **Le développement en libre accès de reproductions « simulées » de la base principale du SNDS.** Ce type d'outils reprend la structure et les formats de données de la base mère mais en remplaçant toutes les données d'origine par des données fictives de façon à l'anonymiser complètement. Ce type de base fictive permet aux chercheurs de se faire une première idée concrète de l'organisation des données dans la base, de manière plus directe que via la lecture du dictionnaire des données de la base. Cela peut être en particulier très utile pour commencer à réfléchir aux compétences, outils analytiques et d'organisation nécessaires pour construire un projet de recherche, imaginer des appariements possibles avec d'autres bases etc. Le Health data hub a développé plusieurs initiatives pour mettre à disposition des données de synthèse et faciliter la prise en main de la base principale du SNDS (en partenariat notamment avec l'INRIA). Ces outils peuvent alimenter les formations au SNDS avec des outils pratiques ;
- **L'ouverture d'accès ponctuels au SNDS, sur demande, pour la réalisation d'étude de faisabilité de projets de recherche.** La logique serait, pour les chercheurs qui ne peuvent se suffire des données de l'ESND pour tester la faisabilité de leur projet, de leur donner accès, de manière temporaire et contrôlée, à l'ensemble du SNDS pour réaliser une étude de faisabilité de leur projet. Un tel accès pourrait être conditionné à la signature d'une charte rappelant les principes à respecter et les engagements des chercheurs, dont l'élaboration pourrait être menée sous l'égide du comité stratégique des données de santé, en associant étroitement la CNIL ;
- **Le développement de « datamarts » thématiques du SNDS,** sous-produits des données de la base principale du SNDS, qui permettent de restituer de manière agrégée les données, avec un accès et une manipulation facilitée pour les utilisateurs potentiels. Le

développement de ces produits est nécessairement progressif, du fait de la charge de travail occasionnée (réalisation des outils, mais aussi documentation et actualisation dans la durée..). Celui-ci pourrait être confié au Health data hub, en appui de la CNAM. De telles données agrégées sont déjà mises à disposition à travers la cartographie des pathologies et des dépenses de l'assurance maladie, des outils comme scansanté ou les données agrégées d'accès libres mises à disposition par le CépiDc.

Recommandation n°15 Développer des outils de « bac à sable » accessibles librement ou de manière simplifiée, permettant de confirmer la faisabilité de projets d'études ou de recherche sur les données de la base principale du SNDS et, si possible, sur d'autres bases de données de santé.

Enfin, même si la loi de 2019 a significativement fait évoluer la responsabilité de l'assurance maladie pour la mise à disposition des données du SNDS, **la CNAM continuera, même après la mise à disposition d'une copie de la base principale du SNDS sur une autre plateforme, à jouer un rôle important dans le système, qu'il convient de clarifier et de consolider.**

En effet, la CNAM restera le concentrateur des données de la base principale du SNDS et l'opérateur de la chaîne de pseudonymisation du SNDS. Ainsi, les appariements directs nécessiteront toujours une intervention de la CNAM. Il faut à ce titre saluer la rénovation récente de la chaîne de pseudonymisation de la CNAM, en réponse à la demande de la CNIL, qui a permis sa quasi-automatisation. Des investissements informatiques suffisants, notamment avec des unités de traitement graphique (GPU), pour assurer ce rôle devront par ailleurs être assurés, pour garantir une qualité de service satisfaisante.

En revanche, le maintien des accès permanents à la CNAM reste encore incertain. Il conviendra de clarifier rapidement l'architecture cible ; si les accès permanents demeuraient à la CNAM, des investissements informatiques devront être réalisés pour assurer leur exploitation dans des conditions satisfaisantes (absence de coupures régulières comme à l'heure actuelle).

2.2.3.4 La « parallélisation » de l'évaluation scientifique et éthique et de l'examen par la CNIL pour garantir le respect des délais de mise à disposition prévus par le futur règlement EHDS

Dans la dernière version connue de la mission, le projet de règlement EHDS impose à l'organisme responsable de l'accès aux données de santé, sollicité par une demande d'accès à des données de santé, de délivrer ou de refuser l'autorisation de traitement dans un **délai de trois mois** à compter de la réception de la demande. Ce délai ne pourra être prolongé, si nécessaire, de trois mois par l'organisme que « *compte tenu de l'urgence et de la complexité de la demande ainsi que du volume des demandes* ». Alors que le projet de règlement prévoyait dans sa version initiale qu'en l'absence de décision dans le délai imparti, l'autorisation devait être regardée comme tacitement délivrée, cette disposition ne semble pas avoir été conservée, laissant entendre que le « *data permit* » doit être expressément délivré.

En outre, le projet de règlement EHDS prévoit, à ce stade, que le détenteur des données doit mettre les données de santé requises à la disposition de l'organisme responsable de l'accès aux données dans un délai déterminé par ce dernier **qui ne doit pas excéder trois mois**. Seulement « *dans des cas justifiés, comme en cas de demande complexe et lourde* » – dont on suppose, en l'absence de précision dans le règlement, que les critères d'appréciation pourront être définis par

les États membres –, ce délai pourra être prolongé pour une période supplémentaire de trois mois par l'organisme responsable de l'accès aux données.

À cela s'ajoute le fait qu'en application de l'article 46 du projet de règlement EHDS, une fois l'autorisation de traitement accordée, l'organisme responsable de l'accès aux données sera tenu de les mettre à disposition dans un **délai de deux mois après les avoir reçues du détenteur des données**, « *sauf si l'organisme responsable de l'accès aux données indique qu'il mettra à disposition les données dans un délai plus long* ».

Ces délais réglementaires d'accès et de mise à disposition des données apparaissent résolument plus souples que ceux initialement envisagés dans la première version du projet de règlement EHDS.

Les délais actuellement en vigueur pour l'examen des demandes d'autorisation par le CESREES et la CNIL – le premier disposant d'un mois renouvelable une fois et la seconde d'un délai de deux mois, également renouvelable une fois –, qui peuvent en théorie conduire à la délivrance d'une autorisation d'un délai total maximal de six mois, apparaissent compatibles avec les délais envisagés dans le cadre du projet de règlement EHDS. Néanmoins, il convient de garder à l'esprit que, comme il a été rappelé dans la première partie du présent rapport, le délai de délivrance de l'autorisation par la CNIL a pu, en pratique, dépasser significativement les quatre mois réglementaires.

La mission estime indispensable que les acteurs de la procédure réglementaire d'instruction des demandes d'accès aux données de santé anticipent d'ores et déjà un respect rigoureux des délais prévus par le règlement EHDS, voire même ambitionnent de se situer autant que possible en-deçà de ces délais. En effet, même si ce dernier ne devrait entrer en vigueur qu'au plus tôt à l'horizon 2027, l'accélération du traitement des demandes d'accès aux données constitue un enjeu essentiel. Une des voies d'accélération consisterait en une « parallélisation » de l'examen de la pertinence scientifique, de la méthodologie de la recherche et de sa finalité d'intérêt public par le CESREES ou, le cas échéant, par le CSE local de la base de données requêtées (cf. partie 2.2.2), et de l'examen de la conformité du traitement aux exigences de protection des données par la CNIL, et, à l'avenir, par l'organisme responsable de l'accès aux données, plus communément dénommé « HDAB » (« *Health Data Access Body* ») sous l'empire du futur règlement EHDS.

Afin de respecter un délai d'examen de la demande d'accès de trois mois, il semble opportun à la mission d'**expérimenter dès maintenant une instruction en parallèle des demandes d'autorisation par le CESREES et la CNIL**, chacune de ces deux instances devant se concentrer sur le périmètre strict de son examen afin d'éviter toute redondance. Une telle « parallélisation », qui n'a pas vocation à empêcher tout échange entre le CESREES et la CNIL en cours d'examen du projet, conduira bien entendu la CNIL, chaque fois que nécessaire, à réserver la délivrance de l'autorisation jusqu'à ce que le CESREES se soit prononcé sur la demande de traitement. Cette « parallélisation » ne fait pas non plus obstacle à ce que la CNIL prenne d'elle-même l'initiative, suffisamment en amont, d'interpeler directement le CESREES sur la question du caractère d'intérêt public de la recherche sur le fondement de l'article 72 de la loi « informatique et libertés », si elle estime que la demande le nécessite.

Recommandation n°16 Expérimenter la parallélisation de l'examen des demandes d'autorisation de traitement de données de santé par le CESREES (ou, le cas échéant, par le CSE de la base de données requêtée) et par la CNIL, en commençant par les demandes d'appariement avec la base principale du SNDS.

2.3 Refonder et renforcer le positionnement du Health data hub en assumant une logique plus fédérative, pour en faire un véritable pilote opérationnel de l'écosystème

2.3.1 Lever l'obligation de centralisation auprès du Health data hub pour l'extension du catalogue du SNDS

La confusion actuelle sur la notion du « catalogue » du SNDS géré par le HDH appelle une clarification. Une distinction doit ainsi être opérée entre :

- d'une part, le catalogue de bases de données qui ne couvre pas l'ensemble de la population mais qui, en raison de leur intérêt particulier, constituent un complément utile à la base principale du SNDS et gagnent à y être appariées ;
- d'autre part, le (ou les) référentiel(s) des bases de données de santé partageables dans le cadre d'une utilisation secondaire, à des fins de recensement de ces bases mais aussi d'information des chercheurs, par la mise à disposition de métadonnées renseignant le contenu de ces bases.

La mission reconnaît l'intérêt d'un catalogue de collections de données à caractère national ou local qui, sous réserve de leur maturité et de leur conformité aux exigences du RGPD, pourraient gagner en visibilité en étant inscrites au catalogue du SNDS. La preuve en est que le catalogue actuel du SNDS compte dix partenaires, et 25 autres sont pressentis pour le compléter.

Toutefois, la question de l'hébergement systématique par le HDH d'une copie des bases inscrites au catalogue du SNDS, qui est aujourd'hui une obligation réglementaire⁸³, concentre les crispations d'une partie des détenteurs de bases de données (cf partie 1.3.2).

Si la logique partenariale du catalogue qui permet, de l'avis même des responsables des bases déjà versées au catalogue, de faire monter en visibilité les bases concernées et de favoriser leur mobilisation dans le cadre de recherches dont les résultats sont publiés à l'international, a du sens, **la répllication de chacune des bases sur la plateforme technologique du HDH ne saurait continuer à conditionner leur inscription au catalogue du SNDS, car c'est un frein à l'enrichissement de ce catalogue.**

Dès lors, la mission recommande une **modification du III de l'article R. 1461-3 pour faire du stockage et de la mise à disposition par le HDH des données des bases inscrites au catalogue du SNDS une faculté et non plus une obligation** : ce stockage centralisé sur sa plateforme

⁸³ En application du 2° du III de l'article L. 1461-3 du code de la santé publique, et du sixième alinéa du même III.

technologique devra nécessairement procéder d'un partenariat conclu avec le responsable de la base.

L'absence d'hébergement sur la plateforme du HDH d'une copie d'une base ne doit pas, pour autant, interdire l'inscription de la base au catalogue du SNDS, en particulier si le contenu de cette base présente un intérêt particulier en termes de chaînage avec la base principale du SNDS. Le responsable de la base de données devra toutefois s'engager au partage de ses données, en respectant les principes définis sous l'égide du Comité stratégique des données de santé et les standards d'interopérabilité établis par le Health Data Hub en concertation avec l'Agence du numérique en santé (voir *infra*, partie 2.4.2.2).

Un corollaire de cette recommandation pour faciliter l'inscription au catalogue du SNDS consiste à alléger la procédure administrative d'inscription au catalogue du SNDS. Celle-ci est aujourd'hui conditionnée à la publication d'un arrêté ministériel. L'inscription au catalogue gagnerait à être validée par le comité stratégique des données de santé.

Recommandation n°17 Faciliter l'enrichissement du catalogue du SNDS en allégeant le processus d'inscription et en faisant du versement d'une copie d'une base de données sur la plateforme du HDH une faculté et non une obligation, sous réserve du respect des principes de partage et des standards d'interopérabilité par les bases inscrites au catalogue.

2.3.2 Renforcer les missions du Health data hub pour en faire un animateur de la mise en réseau des bases de données de santé et un garant du partage

Dès lors que la question de la plateforme technologique du Health data hub et qu'une logique de moindre centralisation auront été clarifiées, le Health data hub pourra jouer **un rôle renouvelé, avec des missions renforcées et un rôle fédérateur** :

- **Un rôle de garant du partage des données dans l'écosystème**, en anticipation d'un positionnement de HDAB dans le cadre du futur règlement européen. Dans le prolongement du rôle moteur qu'il joue actuellement en particulier dans le cadre des groupes de travail du comité stratégique des données de santé, le Health data hub doit jouer un rôle d'impulsion et d'animation pour la définition de principes de partage de données, tel que l'élaboration de grilles de tarification et de contrats types (cf. partie 2.4.3.2), l'élaboration de règles de gouvernance des Comités scientifiques et éthiques ou encore la négociation de principes de reconnaissance scientifique des producteurs de données (cf. partie 2.4.3.3). Le Health data hub doit également jouer un rôle majeur de mise en visibilité des données de santé (voir partie 2.4.2.1). En anticipation du futur règlement européen, l'absence de respect de ces doctrines pourrait être signalée au Health data hub, qui pourrait être doté d'un pouvoir de contrôle. Ceci suppose toutefois une évolution de ses compétences législatives.

- **Un rôle de définition de standards et de normes d'interopérabilité, en associant l'agence du numérique en santé et la délégation du numérique en santé, en particulier concernant les entrepôts de données de santé (voir partie 2.4.2.2)**
- **Un rôle moteur dans l'élaboration de doctrines juridique ou techniques**, à travers le recueil des attentes des utilisateurs et l'animation de travaux de réalisation de doctrines (référentiels et méthodologies de référence, mais aussi doctrines techniques comme sur le principe de pseudonymisation). Ces doctrines seraient validées par le Comité stratégique des données de santé, et, pour les doctrines juridiques, après avis de la CNIL (cf point 2.2.1.2) ;
- **Une mission renforcée d'accompagnement des porteurs de projet** dans l'utilisation des données de la base principale du SNDS mais aussi **dans la mise en œuvre des démarches réglementaires et, le cas échéant, leur mise en conformité avec les exigences du RGPD**. Il appartiendra en particulier au HDH, à ce titre, de développer des outils (formations, tutoriels...) permettant aux porteurs de projet non seulement de pleinement s'approprier les procédures réglementaires mais également de mieux comprendre le contenu et les modalités d'utilisation des bases du SNDS.

Le Health data hub conserverait une plateforme technique permettant, outre la mise à disposition de la copie de la base principale du SNDS, la mise à disposition d'espaces techniques d'espaces de travail, pour les projets multicentriques et les projets nécessitant de très fortes capacités informatiques en termes de volume et de performances. Le HDH continuerait à jouer un rôle pour la réalisation d'outils techniques facilitant l'accessibilité des données, à l'instar du projet de concentrateur du NIR, visant à automatiser l'accès aux données de la CNAV en cas d'appariement direct, ou encore à l'instar des travaux sur l'« OMOPisation » du SNDS. Toutefois, ce rôle technique serait moins prévalant qu'actuellement, du fait du développement de missions d'autres natures.

En revanche, le HDH jouant un rôle plus proche de régulateur sur plusieurs aspects, la mission s'interroge sur le maintien du rôle de **financement de projets innovants** requérant l'utilisation de données de santé (prévu au 8° de l'article L. 1462-1 du code de la santé publique). L'exercice de cette mission semble peu compatible avec la perspective d'exercer les fonctions de HDAB, susceptible d'être doté d'un pouvoir de sanction vis-à-vis des producteurs de données. Si cette mission devait être abandonnée, il conviendrait que le HDH soit plus systématiquement associé dans le lancement par d'autres acteurs (notamment ANR et INSERM) d'appels à projet visant à sélectionner des projets de recherche sur les données de santé.

En outre, la mission souligne que l'élargissement des missions du HDH suppose une **composition pluridisciplinaire encore plus importante de ses équipes**, notamment par l'intégration de juristes, de profils de chercheurs et d'épidémiologistes, de médecins universitaires...

Recommandation n°18 Redéfinir et renforcer les missions du HDH afin d'en faire un acteur fédérateur de l'écosystème de l'utilisation secondaire des données de santé

L'exercice de ces nouvelles missions par le Health data hub suppose que le comité stratégique des données de santé, dans un rôle de pilotage stratégique renforcé (cf partie 2.4.1), s'assure de **l'articulation et la cohérence entre les multiples initiatives en matière de structuration de l'écosystème des bases de données**, afin de clarifier le positionnement du HDH au sein de l'écosystème. En particulier, le rôle du Health data hub en matière de référencement des bases de données en santé devra être clarifié (notamment vis-à-vis de l'Inserm) ; le positionnement de la plateforme du Health data hub par rapport à d'autres plateformes en cours de développement, plateforme France Cohortes, plateforme des données de santé en cancérologie par l'INCA et de la plateforme de données en cancérologie⁸⁴ par l'INCA ou encore du collecteur analyseur de données (CAD) génomiques dans le cadre du plan France Médecine génomique 2025, devront également être précisés (voir partie 1.3.3). La mission recommande que le Comité stratégique des données de santé confie un rôle de coordination des initiatives au Health data hub.

Recommandation n°19 Charger le comité stratégique des données de santé de veiller à l'articulation et à la mise en cohérence des différentes initiatives de structuration de l'écosystème des bases de données de santé, en s'appuyant sur le Health data hub pour coordonner les initiatives.

2.3.3 Une évolution institutionnelle plus profonde à anticiper, dans la perspective du règlement européen

L'article 36 du projet de règlement EHDS prévoit la désignation par chaque État membre d'un ou plusieurs **organismes responsables de l'accès aux données de santé (« health data access body » – HDAB)**, chargés de délivrer les autorisations d'accès aux données de santé à des fins d'utilisation secondaire (« data permits »). Les différentes missions que leur confie le projet de règlement EHDS sont détaillées dans l'encadré ci-dessous.

Il est prévu le **maintien d'un mécanisme décentralisé de demande d'accès à des données de santé** lorsqu'est sollicité l'accès à des données d'un détenteur de données unique dans un seul État membre (article 49 du projet de règlement). Dans ce cas, la demande peut être adressée directement au détenteur de données qui l'instruit et, le cas échéant, délivre l'autorisation dans les mêmes conditions que celles applicables à la procédure de traitement des demandes d'accès par le HDAB⁸⁵.

⁸⁴ Dont l'autorisation a été renouvelée par la délibération n° 2023-040 du 20 avril 2023 de la CNIL.

⁸⁵ Toutefois, dans un délai de trois mois, le détenteur de données devra avoir informé le HDAB, par voie électronique, de toutes les demandes d'accès aux données déposées, de toutes les autorisations de traitement de données délivrées et des demandes de données satisfaites, afin de permettre notamment au HDAB d'assurer sa mission de transparence à l'égard du public sur le traitement des demandes d'accès et les « data permits » délivrés.

Les principales missions des organismes responsables de l'accès aux données de santé (HDAB)

prévue par le projet de règlement EHDS

- l'enregistrement et le traitement des demandes d'accès aux données de santé : à ce titre, les HDAB doivent notamment rendre publics, sur leur site Internet, toutes les autorisations de traitement de données et les demandes d'accès dans un délai de 30 jours ouvrables après la délivrance de l'autorisation de traitement de données ou de la réponse à une demande de données ;
- les opérations de collecte, de combinaison, de préparation et de divulgation des données demandées, y compris lorsque celles-ci sont détenues par d'autres détenteurs de données, et pour lesquelles une autorisation d'utilisation a été délivrée, en vue de permettre leur utilisation secondaire : à cette fin, les HDAB assureront le recueil et la compilation des données de santé nécessaires auprès des différents détenteurs de données, ou donneront accès à ces données, et mettront ces données à la disposition des utilisateurs de données dans un environnement de traitement sécurisé ;
- la perception de redevances au titre de la mise à disposition de données de santé à des fins d'utilisation secondaire ;
- l'envoi gratuit au détenteur de données, avant l'expiration de l'autorisation de traitement de données, d'une copie de l'ensemble de données corrigé, annoté ou enrichi, selon le cas, et une description des opérations effectuées sur l'ensemble de données original ;
- la mise en place d'un système de transparence à l'égard des personnes concernées par l'utilisation de leurs données de santé, les informant de la base juridique sur laquelle repose l'octroi de l'accès aux données, de leurs droits et des modalités d'exercice de ces droits, des mesures techniques et organisationnelles prises pour protéger leurs droits et des résultats des projets pour lesquels leurs données ont été utilisées ;
- la publication d'un catalogue national des ensembles de données, contenant des informations détaillées sur la source et la nature des données de santé électroniques et sur les conditions de mise à disposition de ces données ;
- la publication des sanctions qu'ils prononcent à l'égard des utilisateurs et détenteurs de données qui ne se conforment pas aux exigences du règlement : à ce titre, les HDAB peuvent révoquer l'autorisation de traitement de données et interdire à l'utilisateur de données en infraction tout accès aux données de santé électroniques pendant une période maximale de cinq ans, et, s'agissant des détenteurs de données qui ne s'acquittent pas de leur obligation de partage des données ou ne respectant pas les délais de mise à disposition, leur infliger des amendes pour chaque jour de retard, voire leur interdire de participer à l'EHDS ;
- la facilitation de l'accès transfrontière à des fins d'utilisation secondaire, aux données de santé électroniques hébergées dans d'autres États membres au moyen de DonnéesDeSanté@UE (*HealthData@EU*).

Le règlement EHDS ouvre la possibilité de désigner plusieurs HDAB, sous réserve de désigner un HDAB coordonnateur chargé d'assurer la coordination du traitement des demandes avec les autres HDAB. Dans une telle configuration, il pourrait alors être envisagé de désigner comme HDAB à la fois le HDH et la CNIL, et de répartir entre eux les responsabilités : le HDH, chargé

d'assurer le rôle de HDAB coordonnateur au titre de sa fonction de guichet unique des demandes d'accès aux données de santé, resterait **concentré sur des missions opérationnelles de mises à disposition des données de santé nécessaires à la réalisation des projets de recherche** ; la CNIL conserverait, au titre de sa mission d'agence de protection des données, la responsabilité de **réceptionner les déclarations de conformité aux méthodologies de référence et de délivrer les autorisations d'accès ou les décisions uniques d'autorisation**, et pourrait se voir en outre chargée, dans le cadre de son contrôle exercé *a posteriori* sur les responsables de traitement, de **prononcer les sanctions prévues par le règlement EHDS**.

Toutefois, cette organisation, relativement complexe, tranche avec celle observée dans d'autres pays européens où le choix a d'ores et déjà été fait de **réduire le nombre d'interlocuteurs des utilisateurs de données de santé dans le traitement de leurs demandes d'accès**. C'est en particulier le cas en Finlande où l'organisme FINDATA assure tant l'instruction et la délivrance du « *data permit* » que les opérations de mise à disposition des données requises.

La logique de responsabilisation des responsables de traitement par l'appropriation des procédures simplifiées d'accès aux données de santé plaide, selon la mission, pour confier au seul HDH le rôle de HDAB chargé de centraliser les déclarations de conformité aux méthodologies de référence et de les rendre publics, recentrant ainsi la CNIL pour l'essentiel dans un contrôle *a posteriori* des responsables de traitement. Toutefois, en cas de demande d'accès n'entrant pas dans le champ des méthodologies de référence et nécessitant une autorisation ou, le cas échéant, une décision unique d'autorisation au vu des spécificités de la demande (complexité, sensibilité des données, volumétrie...), le HDH interviendrait en tant que HDAB après délivrance de l'autorisation requise - ou avis conforme - par la CNIL.

S'agissant de la mise en œuvre des sanctions prévues par le règlement EHDS, la circonstance que le HDH soit l'autorité accordant l'autorisation d'accès le rend légitime à la retirer dans le cas où une méconnaissance des exigences réglementaires par l'utilisateur des données serait constatée, le cas échéant à la suite d'un contrôle effectué par la CNIL, à laquelle il reviendrait d'informer systématiquement le HDH des irrégularités relevées à l'occasion de ses contrôles. Le HDH apparaît également le mieux positionné pour sanctionner les détenteurs de données qui ne s'acquitteraient pas de leur obligation de mise à disposition des données demandées dans les délais réglementaires.

Recommandation n°20 Désigner le HDH comme HDAB au sens du règlement EHDS, tout en prévoyant que l'octroi par le HDAB des autorisations pour les accès ne relevant pas d'un référentiel ou d'une méthodologie de référence est accordé par le HDH après avis conforme de la CNIL.

2.4 Réunir les conditions d'une utilisation secondaire optimale des données de santé

2.4.1 Conforter le pilotage stratégique interministériel

Installé en 2019 auprès du ministre de la santé, le comité stratégique des données de santé (Costrat) constitue actuellement une instance utile pour le partage d'informations entre ses

nombreux membres⁸⁶. Les différents groupes de travail constitués sous son égide, avec le pilotage actif du Health data hub, ont aussi permis une concertation appréciée des acteurs et la production de travaux favorisant le partage des données de santé (socle de données des établissements de santé, grille tarifaire, règles de gouvernance ..). Son absence de présidence effective par le ministre de la santé témoigne toutefois d'un rôle centré sur l'animation et le partage, plutôt que d'un rôle stratégique et politique.

Or, malgré le rôle utile joué par le comité, il manque actuellement une instance stratégique pilotant la politique d'utilisation secondaire des données de santé, à même de définir un cadre réglementaire, des principes directeurs et des cadres d'interopérabilité s'imposant aux acteurs, pour faire progresser l'utilisation secondaire des données de santé.

L'absence de véritable gouvernance interministérielle, associant en particulier ministère de la santé et ministère de la recherche, est également très préjudiciable, conduisant à un manque d'articulation entre les initiatives publiques, par exemple en termes de référencement des bases de données (cf point 1.3.3), ou encore dans l'articulation entre les différentes plateformes en cours de constitution (France cohortes et le HDH). De fait, les travaux aujourd'hui menés par les groupes de travail associent des acteurs de la recherche, mais leurs productions n'ont pas vocation à s'appliquer à ces derniers. Par exemple, les règles de tarification ou les règles de gouvernance des comités scientifiques et éthiques ont été définies pour les entrepôts de données de santé hospitaliers, alors qu'il serait souhaitable d'adopter des règles communes à l'ensemble des producteurs de données, par souci de lisibilité et de cohérence pour les utilisateurs de données.

C'est pourquoi la mission recommande de faire évoluer la gouvernance et les missions du comité stratégique des données de santé, pour en faire une instance d'impulsion d'une politique de réutilisation des données de santé, tout en renforçant son rôle opérationnel :

- La présidence du comité stratégique des données de santé serait assurée en commune par le ministère de la santé et le ministère de la recherche, avec au moins une réunion par an avec présidence effective des ministres chargés respectivement de la santé et de la recherche. Le ministre chargé du numérique pourrait être invité à cette occasion, et la direction générale des entreprises serait représentée au sein du comité stratégique des données de santé. Dans une approche « One health », il serait également souhaitable d'intégrer l'agence nationale de sécurité sanitaire et alimentaire nationale (ANSES) au comité, pour mieux intégrer les enjeux intersectoriels avec l'environnement et l'agriculture;
- Un comité restreint serait créé au sein du Comité stratégique, pour jouer un rôle d'instance de pilotage du comité stratégique. Le comité plénier viserait la pleine association des acteurs, le comité restreint la prise de décision sur les orientations stratégiques ;
- De nouvelles missions seraient confiées au Comité stratégique rénové, lui confiant un rôle de définition du cadre de l'utilisation des données de santé (élaboration du « plan d'occupation des sols ») :

⁸⁶ Le Costrat est composé de 25 membres, dont un grand nombre d'administrations et agences sanitaires, 4 représentants des entrepôts de données de santé et deux représentants des associations de patients.

- proposer au ministre des projets de méthodologies de référence et de référentiels (avec un avis de la CNIL) ;
- élaborer des lignes directrices et des références s'imposant aux acteurs pour le partage des données de santé (grilles de tarification et contrats-type ; principes communs de reconnaissance scientifique ; indicateurs de délais, ...) ;
- assurer la cohérence des démarches ministérielles et l'articulation entre les différentes initiatives. A ce titre, le comité stratégique des données de santé serait compétent pour l'ensemble des données de santé, au sens large (réutilisation de données de soins, de données médico-administratives comme de données de recherche).

Le Health data hub continuerait à assurer le secrétariat du Comité stratégique des données de santé mais il sera également nécessaire, compte tenu de l'importance renforcée du comité stratégique et du renforcement de ses missions opérationnelles, que la direction de la Recherche, des Études, de l'Évaluation et des Statistique (DREES) - en lien avec la délégation du numérique en santé (DNS) - et la direction générale de la recherche et de l'innovation (DGRI) renforcent les ressources dédiées à l'animation du Comité.

Recommandation n°21 Rénover la gouvernance et les missions du comité stratégique des données de santé pour en faire une instance interministérielle d'impulsion de la politique d'utilisation secondaire des données de santé ; renforcer les ressources dédiées à l'animation du comité.

Dans la même perspective de renforcer la dimension interministérielle de cette politique et de consacrer le rôle fédérateur du Health data hub dans l'écosystème de la recherche à partir des données de santé, la mission recommande de consacrer sa double tutelle par le ministre chargé de la santé et le ministre chargé de la recherche.

2.4.2 Renforcer la visibilité sur les bases de données et leur interopérabilité

2.4.2.1 Mettre en visibilité les bases de données en santé existantes

On a vu précédemment que le paysage français se caractérisait par un manque de lisibilité sur le contenu des bases de données en santé (variables, étendue, complétude, voire métadonnées) au niveau de chaque producteur de données comme à une échelle nationale. Il est ainsi nécessaire de donner une visibilité opérationnelle plus forte sur les bases de données de santé aux communautés de recherche, académiques et industrielles, à plusieurs niveaux :

- *Au niveau global*, une cartographie transverse, dynamique et actualisée permettant de recenser les bases de données de la façon la plus exhaustive possible sur un périmètre large de bases de données (données de santé issues du soin comme de la recherche). Cette cartographie délivrerait une première information et préciserait les modalités d'accès aux données, avec en particulier un point de contact. Le portail France Épidémiologie constitue un bon exemple d'une telle cartographie, même si celui-ci n'est plus actualisé, que son périmètre ne couvre pas l'ensemble des données de santé et qu'il présente des limites en termes de requêtage ;

- A un niveau plus fin, un (ou des) référentiel(s) plus précis des bases de données, présentant les modèles de données (variables, normes, etc.), une description des métadonnées et des informations permettant de qualifier les données (complétude des variables, périmètre des populations, etc.). Ce référentiel pourrait à termes intégrer des interfaces de programmations (API) pour requêter la présentation des données. La cartographie globale serait articulée avec ce ou ces référentiels plus détaillés ;
- De manière complémentaire, la mise à disposition par les producteurs de données de jeux de données permettant d'illustrer concrètement le contenu des bases de données est également plébiscitée. Ces jeux de données (pouvant être des données simulées) pourraient être intégrés au référentiel de métadonnées.

Une dynamique est en cours sur ce plan et il faut la saluer. Mais il y a plusieurs projets de répertoires ou de référencement des bases de données, sans articulation claire, ces différentes initiatives pouvant être perçues comme redondantes voire concurrentes. Au-delà des initiatives nationales précédemment évoquées (projet FRESH mené par l'INSERM, travaux du Health data hub pour un répertoire public de référencement des bases de données en santé, projet ECLAIRE sur les essais cliniques porté par la DNS), il existe également des projets de référencement de bases de données ou de métadonnées à l'échelle locale. En Ile-de-France, par exemple, l'agence régionale de santé réalise actuellement un état des lieux de l'ensemble des données de santé disponibles et de leur qualité, pour son usage interne mais également à destination de l'ensemble des acteurs régionaux. L'AP-HP développe aussi un data catalogue sur ses données, avec l'ambition d'une documentation précise et la mise à disposition de jeux d'exemples de données. Ces initiatives sont positives mais méritent d'être coordonnées et articulées avec les initiatives nationales (par exemple sur la nature des métadonnées mises à disposition).

Plusieurs options existent pour améliorer l'articulation des différents projets de recensement.

S'agissant du premier niveau de référencement, une première option consisterait en une centralisation en un lieu unique, avec un seul acteur responsable. Cette option présente l'avantage d'être cohérente avec la perspective dressée par le règlement EHDS de rendre public, via l'organisme responsable de l'accès aux données de santé (*Health data access body*- HDAB), un catalogue national de l'ensemble des données disponibles, avec des informations sur la source et la nature de ces données ainsi que sur les conditions de leur mise à disposition.

La désignation du Health data hub comme responsable de cette centralisation serait cohérente avec cette perspective, ainsi qu'avec ses missions actuelles⁸⁷. Celui-ci a d'ailleurs réalisé un premier référencement de ce type pour les données du catalogue du SNDS, en s'inspirant de la structure du portail France épidémiologie, et selon un schéma validé par le comité stratégique des données de santé.

⁸⁷ Parmi ses missions figurent la diffusion d'outils et la contribution à l'élaboration de référentiels pour l'échange et l'exploitation des données de santé (cf Arrêté du 29 novembre 2019 portant approbation d'un avenant à la convention constitutive du groupement d'intérêt public « Institut national des données de santé » portant création du groupement d'intérêt public « Plateforme des données de santé »)

Une seconde option consisterait à maintenir plusieurs recensements/ référentiels, avec un périmètre clarifié (nature des bases de données couvertes, taille critique des bases intégrée, etc.), une harmonisation des données recensées et une meilleure articulation. En tout état de cause, s'il y a pluralité de référencements, il convient de développer des dispositifs techniques permettant, d'une part, une seule remontée d'informations par les producteurs de données, afin de limiter les saisines de données par les producteurs de données et réduire l'effort d'actualisation des données, d'autre part, une capacité à ce que les informations présentes sur un référentiel puissent l'être également sur l'autre référentiel (logique de « marque blanche »).

S'agissant du deuxième niveau de référencement (référentiel de métadonnées), si la centralisation d'une description fine des métadonnées de l'ensemble des bases de données par un unique acteur apparaît complexe, la reprise d'un modèle unique de description des métadonnées (dans une logique de « marque blanche ») est nécessaire. Sur ce point, la réalisation d'un prototype de référentiel de métadonnées par le Health data hub, intégrant 72 métadonnées, constitue un premier acquis.

Dans tous les cas, il est nécessaire de clarifier les articulations de ces projets avec FRESH. Ce projet pourrait être recentré sur la description des métadonnées (2^{ème} niveau de référentiel) pour les données de santé produites par la recherche, compte-tenu de la proximité du porteur de projet avec les acteurs de la recherche. Par ailleurs, un suivi régulier du projet mériterait d'être réalisé au niveau du comité stratégique des données de santé, pour garantir l'articulation des différentes initiatives.

La mission recommande ainsi que le comité stratégique s'empare de ce sujet, examine les différentes options et prenne les décisions nécessaires pour définir un cadre national plus affirmé s'agissant des cartographies de bases de données, visant à :

- mieux articuler les initiatives existantes (cf. *supra*) ;
- définir le contenu de chaque niveau de cartographie ;
- harmoniser les différentes métadonnées produites par les producteurs de données (en cohérence avec les cadres de métadonnées actuellement développés à l'échelle européenne) ;
- définir la fréquence et les mécanismes d'actualisation des données ;
- définir des mécanismes d'incitation aux producteurs de données (et le cas échéant de sanctions).

Ce cadre national devrait être assorti d'engagements communs des producteurs de données :

- la mise en visibilité et mise à disposition des éléments permettant les deux niveaux de référencement ;
- une logique de partage des données affichées, avec des indications sur les processus de partage et un engagement de réponses aux sollicitations ;

- une actualisation régulière (selon une fréquence à définir) pour les deux niveaux de référencement ;
- en cas de pluralité d'initiatives nationales de mutualisation, l'articulation, dans une logique de « marque blanche » et l'utilisation des mêmes règles (notamment sur le type de données).

La mission recommande aussi de déployer ce cadre de manière progressive. A court terme (horizon inférieur à six mois), un premier niveau de recensement pourrait être mis en ligne (version bêta). Une clarification des rôles et des articulations entre les projets serait réalisée, tandis que le cadre commun de « marque blanche » de métadonnées (pour le second niveau de recensement) serait fixé. A moyen terme (18 mois), la mise en ligne de cartographies intégrant une description des métadonnées et des jeux d'exemples serait recherchée, ainsi que la création d'API pour requêter la présentation des données au niveau local.

A terme, le recensement des bases de données existantes, avec une description a minima des métadonnées, permettrait aux équipes tierces d'avoir connaissance de ces dernières et de pouvoir formuler une demande d'intérêt auprès du responsable de traitement, favorisant ainsi une dynamique vertueuse de partage de données.

Ce cadre aurait vocation à s'appliquer y compris aux initiatives privées ; sa négociation gagnerait donc à associer des acteurs privés. Le Health data hub aurait pour mission de s'assurer de la bonne association de ces acteurs lors de l'élaboration de ce cadre.

Recommandation n°22 Définir un cadre national pour mettre en visibilité les bases de données de santé, en clarifiant l'articulation entre les différents projets en cours ; en confier le pilotage au comité stratégique des données de santé.

2.4.2.2 Adopter des standards de données pour renforcer l'interopérabilité des bases, en particulier des entrepôts de données de santé

L'adoption de standards de données commun concerne au premier chef les entrepôts de données de santé (EDS) hospitaliers dont l'hétérogénéité des données, l'absence d'utilisation de nomenclatures communes et le manque de lisibilité des structures de données limitent leur réutilisation (voir partie 1.2.1.3 et annexe 1).

Sous l'égide du comité stratégique des données de santé, un groupe de travail a élaboré en 2023 un socle de données communes pour les EDS hospitaliers. 51 items ont été identifiés, notamment des données socio-démographiques des patients, des données sur l'exposition médicamenteuse, des résultats d'examens de biologie médicales, des données issues de l'examen clinique des patients. Il est prévu que ce premier travail soit prolongé par un second groupe de travail pour favoriser la standardisation de ces données. L'objectif de ce second groupe de travail est de produire une version technique du socle commun avec, pour chaque item, la liste exhaustive des variables attendues, des éléments liés à la standardisation de chaque variable (terminologie, format, présence dans le modèle commun), des métriques de qualité et des éléments de documentation.

La mission recommande que ces travaux de standardisation se poursuivent par la publication de normes d'interopérabilité contraignantes pour les principaux types de variables des entrepôts de données.

Le référentiel d'interopérabilité gagnerait à intégrer une exigence de partage systématique des schémas de structuration des données par les éditeurs de logiciels, afin de faciliter la constitution et l'alimentation des EDS, et plus largement l'interopérabilité entre les systèmes d'information (y compris avec les logiciels de ville). Le cadre juridique récemment complété (articles L 1470-5 et 1470-6 du code de la santé publique) fournit la base juridique permettant une telle obligation, qui devrait être discutée dans ses modalités avec les éditeurs de logiciels et leurs représentants. Une attention particulière devra être portée à l'application de ces normes d'interopérabilité ; sur ce point, les enseignements tirés par la DNS sur le déploiement de normes pour les données primaires (Mon espace santé, Ségur du numérique...) par les éditeurs de logiciel seront précieux.

Il devra également poser les premiers jalons pour permettre à moyen termes la possibilité d'API pour requêter les EDS hospitaliers.

Sur un plan technique, en termes d'architecture de partage des données, et contrairement à ce qui se pratique dans les secteurs numériques, le secteur de la santé ne fonctionne pas par la mise à disposition des jeux de données de manière décentralisée sous forme de « micro-services » avec la publication d'API. En conséquence, les données sont copiées à répétition à de multiples endroits, cela ralentit leur accès et cela pose la question du contrôle de leur circulation et de leurs stockages.

A l'instar des travaux de standardisation pour l'échange de données médicales en cours de déploiement dans le monde, en particulier ceux liés à l'initiative FHIR (Fast Healthcare Interoperability Resources) de l'organisation internationale HL7 et portés en France par l'Agence du numérique en santé, il apparaît particulièrement critique de créer une dynamique de partage fondée sur le recours à de tels standards et aux API correspondantes, comme l'API SMART/HL7 FHIR Bulk Data Access. C'est un enjeu d'efficacité opérationnelle et aussi de passage à l'échelle et d'attractivité internationale du partage de données de santé en France. Pour structurer la dynamique, tous les appels à projets et financements publics pourraient ainsi être conditionnés par le respect des standards communs.

Sur un plan opérationnel, le Health data hub piloterait ces travaux de standardisation des données, dans le cadre de missions renforcées sur ce volet (cf. partie 2.3.2). L'agence du numérique en santé et la délégation du numérique en santé devront être associées étroitement à ces travaux.

Recommandation n°23 Définir, sous l'égide du Health data hub, en associant étroitement l'agence du numérique en santé et en concertation avec les acteurs, des normes d'interopérabilité pour les entrepôts de données hospitaliers ; intégrer dans ces normes une exigence de partage des schémas de structuration des données par les éditeurs de logiciels.

De manière générale, au-delà des EDS, la réutilisation des données de santé sera grandement facilitée par l'utilisation de standards de données et de nomenclatures communes. L'utilisation des standards internationaux (sémantiques comme techniques), en privilégiant les standards ouverts, doit en effet être recherchée. A titre d'exemples, l'utilisation des nomenclatures médicales internationales dans les registres n'est pas systématique.

La mission recommande qu'une réflexion soit engagée sur l'élaboration d'un index sur la qualité des bases de données en santé, afin de favoriser l'utilisation de nomenclatures communes et donner une plus grande visibilité sur le contenu des données (complétude, périmètre, etc.). Celle-ci pourrait s'appuyer sur l'expérience du comité d'évaluation des registres, qui a évalué 64 registres dont 33 registres de cancers depuis sa création. Sur le champ des entrepôts de données de santé hospitaliers, l'AP-HP participe à des travaux internationaux pour établir un modèle de maturité des EDS, et a initié une première auto-évaluation selon la grille d'analyse de ce modèle de maturité comprenant six catégories d'évaluation.

Pour chaque nature de bases de données, des grilles d'évaluation pourraient être réalisées pour mesurer de manière quantitative la représentativité des données (pourcentage de données manquantes globales, ou par variable, taille des populations, ...) et analyser de manière qualitative leur qualité intrinsèque (rareté des données...). Leur élaboration pourrait être pilotée par le Health data hub et validée par le comité stratégique des données de santé.

2.4.3 Consolider le modèle économique et garantir une reconnaissance scientifique pour encourager au partage

2.4.3.1 Favoriser les financements pérennes des bases de données en santé et rendre payant à terme l'accès aux données du SNDS

La fragilité du mode de financement d'une partie des bases de données en santé liée à la prépondérance des appels d'offres peut avoir pour conséquence, une logique contre-productive en termes de partage des données. Face à l'exigence de valorisation économique, les producteurs de données sont incités à limiter ou conditionner l'accès aux données dont ils disposent, d'une part pour réserver une forme d'« avantage compétitif » aux cliniciens ou aux chercheurs de leur structure, d'autre part, pour préserver un potentiel « maximum » de valorisation financière. Cette logique est profondément contre-productive, elle a des conséquences négatives en termes de production de valeur scientifique et probablement en termes de valorisation économique. Il faut donc définir des mécanismes, y compris de financement, de nature à orienter l'ensemble des acteurs, et en particulier les structures qui construisent et exploitent des infrastructures de données de santé, vers le partage de données.

La mission recommande ainsi de favoriser les financements pérennes et suffisants des bases de données, selon plusieurs axes :

- Des financements structurels des entrepôts de données de santé hospitaliers devraient être établis. Le montant des financements doit tenir compte des travaux établis avec le HDH pour évaluer financièrement les coûts de mise en place et d'exploitation des entrepôts de données de santé (cf. annexe 1 sur les entrepôts de données de santé). Les financements devraient tenir compte du volume de données, de leur utilisation (nombre de projets, publications, etc.) et naturellement de leur degré de partage. Les entrepôts financés devront s'engager en contrepartie au respect des normes établies par le comité stratégique des données de santé (interopérabilité mais aussi en matière de fonctionnement des comités éthiques et scientifiques) ;

- De même, un financement pérenne du Health Data Hub doit également être intégré dans l'objectif national des dépenses d'assurance maladie (ONDAM), en substitution des financements par le biais du fonds de transformation de l'action publique (jusqu'en 2023) ;
- Les financements des cohortes devrait également être pérennisés et établis de manière pluriannuelle. La stratégie nationale pour les cohortes proposée par l'INSERM prévoit la création d'un comité des cohortes avec un financement annuel des cohortes via des appels d'offre par l'agence nationale de la recherche) ;
- La consolidation des financements des registres devrait être étudiée, dans le prolongement du rapport du Haut Conseil à la santé publique de 2021. Les registres sont actuellement financés par les agences sanitaires (Santé publique France et l'INCA, notamment pour les registres cancer), par les établissements de santé (via une dotation Mission d'intérêt général) pour les registres qui leur sont adossés, ou encore par des financements divers (agences régionales de santé, collectivités territoriales, appels à projets...). Le rapport de 2021 préconisait le financement par un guichet unique pour mieux répartir les financements selon une stratégie nationale, les financements étant conditionnés au respect de cahiers des charges et l'obtention d'une labellisation. Aucune suite n'ayant été donnée à ce jour à ces travaux, une nouvelle réflexion mériterait d'être engagée sur ce point dans le cadre de la future feuille de route de l'utilisation secondaire des données de santé.

Il conviendra également de veiller à ce que l'ensemble de ces financements encouragent les producteurs de données à partager les données produites.

A l'inverse, l'accès aux données de la base principale du SNDS est aujourd'hui gratuit, alors que de nombreux acteurs seraient prêts à financer cet accès, à condition que les délais d'accès soient satisfaisants. Dès lors que ce sera le cas, une tarification des données du SNDS gagnerait à être mise en place, comme c'est actuellement le cas pour l'accès aux données du PMSI via l'ATIH. Cette dernière prévoit aujourd'hui un accès payant pour les bureaux d'étude et laboratoires de recherche avec un système de redevances⁸⁸ liées à la préparation, la mise en qualité des données et leur mise à disposition. Une tarification majorée des données de la base principale du SNDS pourrait être établie en cas de chaînage des données. Cette recommandation suppose une évolution d'ordre législatif (article L 1461-2 du CSP).

Recommandation n°24 Établir des financements pérennes et suffisants des infrastructures de données de santé (entrepôts de données de santé hospitaliers, registres, cohortes) et du Health data hub pour favoriser la logique de partage des données.

Recommandation n°25 Préparer la mise en place un accès payant aux données à la base principale du SNDS dès que les délais d'accès le permettront

2.4.3.2 Etablir un corpus de règles concernant la valorisation des données

On a vu précédemment que des projets de réutilisation de données pouvaient être abandonnés ou significativement retardés par l'absence de lisibilité sur les règles de valorisation des données,

⁸⁸ Redevance forfaitaire de 500 euros pour chaque base de données demandée et redevance de 34 centimes par ligne par fichier demandé (cf. délibération CA ATIH n°5 du 11 mars 2021).

par la grande variabilité dans les modalités de tarification, et par des temps de négociations importants.

De premiers travaux pour un modèle commun de tarification ont été menés par le comité stratégique des données de santé, au sein d'un groupe de travail dédié. Dans ce cadre, deux grilles de tarification des données ont été partagées par deux entrepôts de données de santé hospitaliers (AP-HP pour la Banque nationale de données des maladies rares et GCS HUGO). Ces grilles permettent d'identifier de manière très fine les principaux postes de coûts liés à la mise à disposition de données à un porteur de projet, afin de générer une tarification sur la base de ces coûts en accord avec le règlement européen sur la gouvernance européenne des données (*Data government act*). Le calcul des coûts repose sur la mesure du temps homme nécessaire à la contractualisation, à chaque étape du projet (examen de la faisabilité du projet et du devis, dossier administratif et réglementaire, etc.), et sur l'application de taux jours homme en fonction d'une grille de salaires commune (variant selon les profils des ressources concernées). Ces deux grilles ont fait l'objet d'une première phase de test par cinq entrepôts au premier semestre 2023. Cette phase de test a été prolongée à l'automne 2023 pour six mois, « afin de consolider ces premières observations et d'obtenir des métriques plus précises sur la réalité de la tarification entre dépositaires de données et utilisateurs ».⁸⁹

Si la mission salue la réalisation de ces premiers travaux, elle recommande leur approfondissement et surtout leur accélération, avec deux options dans la mise en œuvre opérationnelle de cette recommandation.

Une première option consisterait à établir une grille de tarif unique pour les redevances, qui serait adoptée par le pouvoir réglementaire et rendue opposable pour l'ensemble des producteurs de données. La Finlande constitue à ce titre avec Findata un modèle intéressant, avec une grille tarifaire publique, adoptée par décret, décomposant la tarification :

- du permis d'accès aux données,
- des coûts d'extraction et de mise à disposition des données par le producteur de données,
- des coûts liés au processus de pseudonymisation et d'anonymisation des données (avec l'application d'un financement horaire forfaitaire) ;
- et des coûts d'accès à distance à l'environnement sécurisé.

Le choix d'un modèle proche conduirait à réorienter les travaux du comité stratégique des données de santé pour retenir un modèle unique de tarification. Celui-ci pourrait s'appuyer sur les deux exemples de grilles en cours de test, en visant autant que possible une simplification des modalités de calcul, avec des montants forfaitaires, à l'image du modèle finlandais. Ce modèle de tarification aurait pour ambition de s'appliquer à l'ensemble des producteurs de données, au-delà des entrepôts de données de santé hospitaliers actuellement visés par la phase de test du groupe de travail du comité stratégique. Il sera donc essentiel d'associer à son élaboration de manière

⁸⁹ Il est notamment prévu l'utilisation en conditions réelles des deux outils d'estimation des coûts d'un projet au sein des organismes partenaires, en mesurant le nombre et le type de projets et d'études déposés auprès des organismes partenaires ; le nombre de devis réalisés et les montants de ces devis par type de projet ; et le temps consacré à l'estimation des coûts et à la génération des devis par les équipes des organismes partenaires).

plus importante qu'actuellement des représentants des cohortes (notamment Inserm Transfert) et des registres, mais aussi des représentants des universités (par exemple à travers les sociétés d'accélération du transfert des technologies-SATT).

Si cette option est retenue, la mission recommande que cette grille de tarification unique intègre une modulation des tarifs en fonction du profil du demandeur (universitaire, PME, grande entreprise...), ce qui est compatible avec le cadre juridique européen actuel (voir encadré *infra*). Elle préconise aussi d'inclure les paiements forfaitaires pour bénéficier de l'expertise du producteur de données, pour qualifier les données, les annoter, ou encore les analyser, en contrepartie de reconnaissances scientifiques.

L'adoption d'une grille tarifaire unique aurait pour avantage de renforcer sensiblement la lisibilité du système, et de permettre une accélération notable du temps de mise à disposition des données. Ses principaux inconvénients résident dans son adaptabilité limitée aux caractéristiques des différents producteurs de données, et potentiellement à un risque de minoration des recettes financières, par rapport à un système de négociations de gré à gré.

Une seconde option consisterait à maintenir l'existence de plusieurs grilles de tarification de référence (deux à trois), sous l'égide du comité stratégique des données de santé, parmi lesquelles les acteurs pourraient choisir.

Cette seconde option permettrait un aboutissement à brève échéance (3 mois) de l'adoption de grilles de référence, à condition, comme dans l'option précédente, d'associer de manière plus étroite les acteurs autres que les EDS hospitaliers à l'élaboration de ces grilles dans la phase de test et de validation des grilles.

Cette option présenterait aussi l'avantage de laisser une plus grande souplesse aux producteurs de données. Elle présente toutefois l'inconvénient d'une moindre lisibilité et simplicité pour l'accès aux données.

En tout état de cause, quelle que soit l'option qui serait retenue, une obligation de transparence des producteurs de données devra être définie en matière de tarification (obligation de publication sur le site du producteur de données du montant des redevances), et plus globalement en matière de processus d'accès aux données, sans attendre l'entrée en vigueur du règlement européen EHDS.

Recommandation n°26 Approfondir et accélérer les travaux pour établir des grilles tarifaires harmonisées pour le financement par redevances des bases de données en santé, et instaurer une obligation de transparence sur les modalités de tarification des producteurs de données.

Un cadre juridique européen de la valorisation économique des données qui reste incertain

Le règlement européen sur la gouvernance européenne des données, entré en vigueur en juin 2022 et applicable depuis septembre 2023, encadre les redevances pouvant être perçues par des producteurs de données, qui doivent être basées sur les coûts liés au traitement des demandes de réutilisation des données- (article 6 du règlement) :

- coûts de reproduction, de fourniture et de diffusion des données ;
- coûts induits par l'anonymisation ou les autres formes de préparation des données personnelles ou confidentielles ;
- coûts induits par la maintenance de l'environnement de traitement sécurisé ;
- coûts d'acquisition des droits des tiers en dehors du secteur public
- coûts d'assistance aux réutilisateurs cherchant obtenir le consentement des personnes physiques ou des ayants droit.

Cette approche d'une tarification au coût marginal va à l'encontre d'une tarification basée sur les coûts complets, en excluant notamment la facturation des coûts de collecte et de production.

L'article 6.4 du règlement DGA permet toutefois d'appliquer des tarifs réduits (pouvant aller jusqu'à la gratuité) pour certains types d'acteurs comme les petites et moyennes entreprises, les startups, les demandes de la société civile et les demandes des établissements d'éducation. Il permet aussi la tarification de services complémentaires.

Toutefois, aucune jurisprudence n'est encore venue préciser la portée de cette définition de la redevance. De plus, les négociations en cours sur le règlement d'espace européen des données de santé pourraient conduire à des ajustements de cette règle pour les données de santé. Une option, en discussion dans le cadre de la négociation sur le règlement européen des données de santé, est de permettre que les redevances incluent tout ou partie des coûts de mise en place de l'infrastructure nécessaire à la collecte et à la production des données. La mission est favorable à une telle option, la logique des coûts marginaux réduits à la mise à disposition des données ne représentant qu'une part mineure des coûts nécessaires pour obtenir et partager des données de qualité.

Par ailleurs, l'étape de contractualisation est trop souvent longue et fastidieuse (au-delà de la négociation sur le montant des redevances). C'est notamment le cas pour les négociations entre acteurs publics.

La mission recommande l'élaboration de contrats-types pour l'accès aux données (à l'instar de la convention unique en matière de recherche clinique⁹⁰), en commençant par deux cas de figure :

- le cas du financement par redevances (en intégrant la facturation des « services associés » qui peuvent être facturés dans le cadre du règlement européen *Data government act*) ;
- l'examen de la situation des projets entre deux (ou plusieurs) acteurs publics, pour lesquels d'autres logiques que des contrats pourraient être imaginées, comme le recours à des

⁹⁰ Arrêté du 28 mars 2022 fixant le modèle de convention unique prévu à l'article R. 1121-3-1 du code de la santé publique

clauses générales d'utilisation, qui pourrait plus particulièrement être étudié pour les conventions avec la CNAM pour l'accès aux données de la base principale du SNDS.

Ces contrats-type gagneraient à fixer des délais maximums de négociation.

Concernant les autres modèles de valorisation qui coexistent avec le paiement de redevance (cf encadré *ci-après*), des travaux supplémentaires devront être engagés, en veillant à leur compatibilité avec le cadre juridique européen. Ces travaux devront associer l'agence de l'innovation en santé ainsi que les acteurs économiques (par exemple dans le cadre du comité stratégique de filière Industrie et santé, qui a déjà conduit de premiers travaux sur cette thématique en 2020⁹¹). Ceux-ci viseront à définir des références communes et à établir des contrats-types.

Recommandation n°27 Définir des modèles de contrats-type pour l'accès aux données de santé, en commençant par le modèle de tarification par redevances.

Les modèles de valorisation co-existant avec le paiement de redevances

- Une participation de l'institution du producteur de données à la propriété intellectuelle générée. Un tel modèle peut être négocié avec les entreprises de grande taille (dépôts de brevets communs assez fréquents et licence exclusive sur un ou plusieurs champs d'application). Il est à l'inverse peu adapté pour les partenariats avec les startups en raison de frilosités des investisseurs vis-à-vis de tels montages ;
- Le versement lors de la mise sur le marché à l'établissement du producteur de données d'une rémunération adossée au chiffre d'affaires de l'entreprise (royalties), dont le montant est le plus souvent majoré. Une alternative consiste en une entrée au capital de l'entreprise. De tels modèles peuvent être développés avec des petites entreprises. Toutefois, l'appréciation de l'apport des données dans l'innovation peut être complexe à mesurer, ce qui implique d'étudier d'entrée de jeu l'éventualité de cette option dans son principe, sans toutefois la définir au préalable dans ses modalités concrètes.
- Une participation au capital (pour les petites entreprises). Cette configuration beaucoup plus rare relève de transfert de technologie développée dans le cadre d'une recherche par une université, un organisme national de recherche ou un hôpital.

2.4.3.3 Définir des lignes directrices pour garantir une reconnaissance scientifique des producteurs de données

Afin de rassurer les producteurs de données et d'encourager le partage des données, le Comité stratégique des données de santé pourrait utilement définir des lignes directrices visant à garantir la reconnaissance scientifique du travail des équipes qui ont œuvré à la constitution et à

⁹¹ « Recommandations pour un modèle économique de la mise à disposition des données de santé issues du Système national des données de santé » – comité stratégique de filière des industries et technologies de santé – groupe de travail sur la valorisation des données partagées et le modèle économique de la plateforme Health Data hub, février 2020

l'alimentation d'une base de données. À cet effet, le Comité stratégique des données de santé pourrait définir des lignes directrices sur :

- les conditions de la valorisation des producteurs de la base dans le cadre des publications scientifiques tirées de son exploitation par des utilisateurs extérieurs ;
- le principe et les modalités d'un embargo d'une durée acceptable - quelques mois - qui garantisse aux producteurs de la base, notamment ceux qui accepteraient l'inscription de celle-ci au catalogue du SNDS, l'exclusivité pendant cette période de l'utilisation de leurs données croisées avec celles de la base principale du SNDS.

La définition de ces principes, qui devront être compatibles avec les recommandations internationales, pourra s'appuyer sur les travaux déjà menés par le Health data hub.

Recommandation n°28 Etablir des principes de reconnaissance scientifique pour la réutilisation de données (principe et durée d'un embargo, règles de publication, ...) pour favoriser le partage

2.4.4 Renforcer la transparence vis-à-vis des patients sur l'utilisation secondaire de leurs données de santé et faciliter l'exercice des droits

2.4.4.1 Tendre vers un portail de transparence centralisé

La **multiplication des portails de transparence** mis en place par les organismes détenteurs et utilisateurs de données de santé ne facilite pas la lisibilité pour les citoyens sur l'usage actualisé qui est fait de leurs données.

Dans un souci de simplification de l'accès des personnes concernées aux informations concernant les traitements réalisés à partir de leurs données, la mission préconise la **mise en place d'un portail de transparence qui serait centralisé par le HDH** et sur lequel seraient versés par chaque organisme détenteur d'une base de données et chaque responsable de traitement les informations que les articles 13 et 14 du RGPD leur imposent de porter à la connaissance du public⁹². Ce portail pourrait également utilement mettre à la disposition du public l'avis du comité d'éthique sur le projet de recherche ainsi que les résultats de l'étude que le porteur de projet s'est engagé à publier.

Ce portail ne concernerait pas les traitements relatifs aux recherches impliquant la personne humaine.

Le déploiement d'un tel portail centralisé s'inscrirait dans le prolongement de la mission qu'assure à l'heure actuelle le HDH en tant que guichet unique de l'ensemble des demandes de traitement de données de santé et au titre de laquelle il a mis en place un **répertoire public des projets de recherche**, lequel reste néanmoins lacunaire en termes d'informations livrées au public. Les

⁹² Identité et coordonnées du responsable de traitement, coordonnées du DPO, base juridique et finalités du traitement, catégories de données concernées, destinataires ou catégories de destinataires des données et, le cas échéant, l'information relative à l'intention du responsable de traitement d'effectuer un transfert de données à caractère personnel à un destinataire situé dans un pays tiers ou une organisation internationale.

détenteurs de bases de données et les porteurs de projets pourraient ainsi, sur leur site Internet, renvoyer les personnes concernées vers le portail du HDH pour la consultation des informations requises par le RGPD. La constitution de ce portail centralisé ne ferait cependant pas obstacle au maintien, sur le site du détenteur de données ou du porteur de projet, de son propre portail de transparence.

Recommandation n°29 Charger le HDH de mettre en place un portail de transparence centralisé recensant les traitements sur les données de santé.

2.4.4.2 Simplifier les conditions d'exercice des droits d'opt-out des patients

La mission recommande la mise en place d'une **plateforme centralisée pour l'exercice par les patients de leurs droits concernant leurs données de santé**, dont leur droit d'opposition, s'agissant de l'ensemble des traitements recourant à leurs données dans le cadre de recherches, d'études et d'évaluations dans le domaine de la santé. Cette plateforme serait gérée par le HDH, lequel aurait déjà l'obligation (en anticipation de ses fonctions de HDAB) de recenser l'ensemble de ces traitements.

Dans le cadre de cette plateforme, le HDH pourrait agir en tant que **mandataire de la personne concernée pour l'exercice de ses droits auprès du responsable de traitement**⁹³. Interrogée sur les conditions de mise en œuvre d'une plateforme centralisée d'exercice des droits, la CNIL a insisté auprès de la mission sur la nécessité :

- de définir la responsabilité et les obligations des responsables de traitement afin que les demandes des personnes soient prises en compte dans les délais prévus par le RGPD ;
- d'informer effectivement et individuellement les personnes de l'existence de cette plateforme et de son champ d'application ;
- de maintenir des modalités alternatives d'exercice des droits afin de garantir leur effectivité et de ne pas délier les responsables de traitement de leurs obligations – la responsabilité finale leur incombant – et de prendre en compte la fracture du numérique ;
- de définir et mettre en place un moteur de recherche ou des filtres afin que les personnes puissent avoir uniquement accès aux informations sur les recherches qui les concernent.

La centralisation de l'exercice des droits, notamment d'opposition, au niveau d'une plateforme gérée par le HDH suppose, bien entendu, que ce dernier dispose d'une vision exhaustive et réactualisée de l'ensemble des traitements utilisant des données de santé (y compris les études réalisées à partir des accès permanents au SNDS). Cette plateforme de centralisation de l'exercice des droits devrait être mise en œuvre en lien avec la CNAM compte tenu de ses responsabilités

⁹³ Il convient de rappeler, à cet égard, qu'en application de l'article 77 du décret n° 2019-536 du 29 mai 2019⁹³, une demande d'exercice des droits (information, accès, rectification, effacement, limitation du traitement, portabilité, opposition) peut être présentée par une personne spécialement mandatée à cet effet par le demandeur, si cette personne justifie de son identité et de l'identité du mandant, de son mandat, ainsi que de la durée et de l'objet précis de celui-ci. Une recommandation de la CNIL en date du 27 mai 2021⁹³ est venue préciser les modalités pratiques d'exercice des droits conférés par le RGPD par le biais de personnes physiques ou morales (les mandataires) mandatées par les personnes souhaitant exercer ces droits auprès des responsables de traitement.

en matière d'exercice des droits⁹⁴ des personnes concernées par l'utilisation de leurs données dans le cadre du SNDS, et de son rôle dans la gestion de la chaîne de pseudonymisation.

Recommandation n°30 Étudier la mise en place d'une plateforme centralisée gérée par le HDH pour l'exercice par les patients de leurs droits concernant leurs données de santé, dont leur droit d'opposition, s'agissant de l'ensemble des traitements recourant à leurs données dans le cadre de recherches, d'études et d'évaluations dans le domaine de la santé.

À terme, la simplification de l'exercice par les patients de leurs droits concernant la réutilisation de leurs données de santé pourrait emprunter la voie de l'outil « **Mon espace santé** », qui pourrait être interconnecté à la plateforme centralisée du HDH d'exercice de ces droits : la personne concernée pourrait ainsi, sur son espace numérique de santé, se voir offrir la possibilité de s'inscrire sur un registre national des oppositions à la réutilisation des données de santé ou de désigner les catégories de recherches – par exemple, « recherches publiques » ou « recherches privées » – pour lesquelles elle souhaite ou ne souhaite pas que ses données soient réutilisées.

2.4.4.3 Envisager à terme la possibilité d'une réutilisation à des fins secondaires des données de Mon espace santé

D'une manière générale, la stricte séparation entre usage primaire et usage secondaire peut d'autant plus être questionnée que le projet de règlement européen EHDS intègre, à son article 33, dans le champ des données devant être mises à disposition à des fins d'utilisation secondaire les **données issues des dossiers médicaux électroniques**⁹⁵ (DME), qui constituent le format européen de collecte et d'échange de données de santé primaires en application des articles 5 et 6 de ce projet de règlement.

Cette perspective européenne de mobilisation à des fins d'usage secondaire de données collectées à titre primaire dans les dossiers médicaux électroniques interroge **l'impossibilité actuelle, en France, d'utiliser les données de santé saisies dans l'espace numérique de santé (ENS) et le dossier médical partagé (DMP)** – composante de l'ENS – à d'autres fins que celles ayant sous-tendu leur collecte, c'est-à-dire « *favoriser la prévention, la coordination, la qualité et la continuité des soins* » conformément au premier alinéa de l'article L. 1111-14 du CSP.

Or les données saisies sur les espaces numériques de santé de chacun des assurés sociaux, dont certaines sont versées directement par ces derniers (poids, taille, consommation d'alcool, de tabac...), constituent une source de données de santé précieuses à fort potentiel en termes d'utilisation secondaire à des fins de recherche. Selon des informations fournies par la délégation du numérique en santé, 7,4 millions de dossiers médicaux partagés (DMP) ont été alimentés par les patients sur les 36 derniers mois à fin octobre 2023, avec 13,3 millions de documents⁹⁶ versés ou alimentés par les patients sur cette période. S'agissant des acteurs de santé, ils ont alimenté,

⁹⁴ Cf. b du II de l'article R. 1461-3 et avant-dernier alinéa du I de l'article R. 1461-9 du code de la santé publique.

⁹⁵ Qui correspondent à l'« ensemble de données de santé électroniques relatives à une personne physique collectées dans le système de santé et traitées à des fins de soins de santé » (point m) du paragraphe 2 de l'article 2 du projet de règlement).

⁹⁶ Les certificats déposés par les patients, les comptes rendus de biologie et les renseignements saisis sur le profil médical « Mon espace santé » constituant le trio des documents les plus alimentés dans « Mon espace santé ».

sur la même période, 45,2 millions de DMP et y ont versé ou alimenté 206,5 millions de documents⁹⁷.

Si la mission est consciente que la priorité reste, pour l'instant, la pleine appropriation de l'outil Mon espace santé par les patients et leurs professionnels de santé à des fins d'utilisation primaire, au service de la coordination des parcours de soins, elle appelle le Gouvernement et le législateur à se saisir, dans un délai pouvant coïncider avec la mise en œuvre du règlement EHDS, de la question de la **levée des restrictions concernant l'utilisation des données de santé saisies sur Mon espace santé à des fins de recherche, d'étude ou d'évaluation dans le domaine de la santé**. À cet égard, la mission relève que le comité citoyen du numérique en santé s'est prononcé en faveur d'une réutilisation à des fins de recherche des données saisies dans l'espace numérique de santé. Par ailleurs, la CNAM gagnerait à anticiper cette possibilité future, notamment en termes de développement informatique pour les données renseignées par les usagers dans leur profil médical comme pour les données structurées saisies par les professionnels de santé.

La mission recommande ainsi qu'à terme, la réutilisation des données saisies dans les espaces numériques de santé à des fins de recherche soit rendue possible **moyennant le consentement exprès de la personne concernée à saisir sur son espace santé**. Ce mécanisme d'« opt-in » pour la réutilisation des données de l'espace numérique de santé serait de nature à garantir la confiance du patient dans la réutilisation de ces données. Afin de mieux familiariser, dès maintenant, les citoyens avec l'importance de leurs données pour l'avancée de la recherche en santé, il pourrait d'ores et déjà être envisagé de mettre en place dans les espaces numériques de santé une information destinée à leurs titulaires sur l'intérêt de la réutilisation des données de santé à des fins de recherche.

Recommandation n°31 Envisager, à terme, la possibilité d'une réutilisation des données saisies dans les espaces numériques de santé à des fins de recherche, moyennant le consentement exprès de la personne concernée qui le saisirait sur l'application "Mon espace santé".

Recommandation n°32 Mettre en place dans les espaces numériques de santé une information destinée à leurs titulaires sur l'intérêt de la réutilisation des données de santé à des fins de recherche.

2.4.5 Développer et consolider certains usages secondaires

2.4.5.1 Compléter la base principale du SNDS et développer certaines bases de données manquantes pour favoriser l'usage secondaire

La complétude des bases de données en santé est propice à l'utilisation secondaire des données de santé, c'est pourquoi l'enrichissement du patrimoine de données français doit constituer un des objectifs de la future feuille de route de réutilisation des données de santé. Il s'agit d'une des

⁹⁷ Parmi lesquels figurent notamment, par ordre : les comptes rendus d'examen biologiques, les prescriptions de médicaments, les comptes rendus ou fiches de consultation ou de visite, les comptes rendus d'imagerie médicale...

missions confiées au comité stratégique des données de santé⁹⁸. Or, si plusieurs projets sont en cours pour consolider le patrimoine de données français (à travers un enrichissement de la base principale du SNDS ou à travers la création de nouvelles bases de données), ces derniers restent peu débattus et ne font pas l'objet d'une programmation stratégique dans le cadre du Comité stratégique.

En premier lieu, l'enrichissement de la base principale du SNDS doit être poursuivi, tant du fait de l'importance stratégique de cette base de données que de ses manques :

- Prévues par la loi, l'intégration au sein de la base principale du SNDS des données des MDPH, ainsi que des bases de données créées pendant le Covid (relatives aux vaccinations et aux tests de dépistage) doit pouvoir être effective en 2024 ;
- Le projet de base de données Labo-SI, pilotée par la direction générale de la santé et la délégation du numérique en santé, est destiné à rassembler des résultats de tests de dépistage en élargissant la logique de la base de données sur les tests de dépistage Covid pendant la crise Covid à d'autres dépistages. Ce projet de création dont la réalisation est prévue en 2024 gagnerait à viser d'emblée l'intégration des données au sein de la base principale du SNDS ;
- De nombreux acteurs de la recherche s'expriment en faveur d'un enrichissement du SNDS par des données relatives aux déterminants de santé des individus, pour faciliter notamment les projets de recherche. La logique annoncée par la DREES est de cibler quelques variables clés, pour éviter la constitution d'une base de données intégrant nativement un trop grand nombre d'informations individuelles (par exemple des informations sur les revenus ou les carrières). Pour cette raison, les appariements « à la source » seraient limités, tandis que les projets de recherche nécessitant des informations complètes relèveraient d'appariements *a posteriori*. Les perspectives privilégiées à ce stade, qui mériteraient d'être partagées en comité stratégique des données de santé et précisées dans leur calendrier comme dans leur trajectoire, sont les suivantes :
 - Une intégration des données contenues dans les certificats de naissance ;
 - Une intégration d'informations géographiques qui ne soient pas trop identifiantes (à l'image des codes IRIS) ;
 - L'intégration des données relatives au poids, à la taille ou aux principaux facteurs de risque comportementaux (consommation d'alcool, de tabac, éléments sur l'activité physique), souhaitée par de nombreux acteurs, mérite d'être davantage étudiée. Plusieurs options sont envisageables :
 - une intégration des données des entrepôts de données de santé hospitaliers ainsi, qu'à termes, une intégration des données des logiciels de médecine de ville. En particulier, un renseignement de ces données lors des rendez-vous de prévention (examens de santé gratuits aux âges de 18 ans, 45 ans, 60 ans et 70 ans, en cours de déploiement par l'assurance maladie) pourrait être recherché ;

⁹⁸ Le comité stratégique des données de santé identifie les catégories de données manquantes et émettre des recommandations en matière de collecte ou de production de ces données auprès des producteurs de données (article 1 de l'arrêté du 29 juin portant création du comité stratégique des données de santé).

- une exploitation des données renseignées par les patients dans Mon espace santé (cf. partie 2.4.4.3).

Recommandation n°33 Élaborer à l'échelle du comité stratégique des données de santé un plan d'enrichissement de la base principale du SNDS et en assurer un suivi régulier.

En second lieu, le développement de nouvelles bases de données doit être encouragé ou conforté sur certains champs. Certains projets s'inscrivent déjà dans cette dynamique, à l'instar du projet d'entrepôt de médecine de ville (cf. encadré partie 1.2.1.4) ou, dans un registre différent, du collecteur d'analyse de données pour les données génomiques ou de la mission de structuration des biobanques. Cette dernière, financée à partir de 2023 dans le cadre de France 2030, vise à structurer le paysage au niveau national, caractérisé par une multitude de centres, mais aussi à soutenir la création de nouvelles collections biologiques à fort potentiel de réutilisation.

D'autres champs méritent la constitution de bases de données, même si cette première analyse mériterait d'être complétée et priorisée par le comité stratégique des données de santé :

- Certaines priorités de santé publique restent très peu couvertes par les bases de données en santé. C'est notamment le cas de la santé mentale ou, dans une moindre mesure, des pathologies chroniques ;
- Certaines zones géographiques sont insuffisamment couvertes par les bases de données en santé, à l'instar des départements et collectivités d'Outre-mer, qui présentent pourtant des enjeux forts en termes de santé publique et de santé environnementale. A cet égard, la mission soutient la proposition de l'Inserm de développer des cohortes spécifiques pour les départements d'outre-mer, telle qu'elle est proposée dans le projet de stratégie sur les cohortes de l'Inserm ;

Recommandation n°34 Poursuivre les appels à projet pour développer des bases de données en santé sur de nouveaux champs définis par le comité stratégique des données de santé, comme la santé mentale, les pathologies chroniques ou portant sur certaines zones géographiques peu couvertes comme les départements et collectivités d'Outre-Mer.

Enfin, même s'il s'agit d'un constat ancien, l'absence de bases de données sur les remboursements de complémentaires santé demeure une lacune. Aucune suite n'a été actuellement donnée à l'expérimentation MONACO⁹⁹. Des échanges mériteraient d'être menés par la DREES avec les complémentaires santé pour clarifier la position publique et les perspectives de création de bases de données (ou d'appariements entre les données d'AMO et d'AMC).

2.4.5.2 Anticiper l'ouverture vers d'autres champs

Comme rappelé précédemment (cf. partie 1.2.2.2), l'intérêt des croisements entre bases de données de santé et la plus-value des analyses multicentriques est évident. Ce constat s'étend naturellement aux croisements de données intersectoriels, qu'il convient de faciliter et d'encourager, en particulier dans une perspective de santé globale (approche *One health*) et de développement de la prévention en santé.

⁹⁹ Cette expérimentation d'appariement entre les données d'assurance maladie et les données des complémentaires santé a été menée en 2012.

Un premier enjeu consiste à connaître (de manière macroscopique) les données existantes sur les autres champs sectoriels que celui de la santé, en disposant d'une visibilité sur la complétude des données, leur qualité et leur interopérabilité.

Cet enjeu est prégnant dans le domaine de l'environnement, où les bases de données ont une maturité particulièrement variable, comme l'illustre le tableau ci-dessous.

Tableau 5 : Niveau de maturité des modèles et données environnementaux disponibles à l'échelle nationale métropolitaine

Niveau de maturité	Très faible	Modéré	Satisfaisant	Très bon
Besoin	Génération de données et modèles	Harmonisation (p. ex. entre villes/région), nettoyage, développement des données existantes	Pérennisation ou affinage des systèmes de surveillance/modèle existants	Existence de modèles pérennes avec une bonne résolution spatiale ou temporelle
Exemples	Usage des pesticides (échelle infra-départementale) Bases de données sur la composition des produits alimentaires transformés	Polluants de l'eau (en nombre limité) Certains polluants de l'air réglementés (benzène...) Bruit ¹⁰⁰ Trafic routier ¹⁰¹	Particules fines, échelle de l'ordre du km (PM _{2,5} , PM ₁₀) (source : Ineris)	Données météorologiques (température, échelle grossière) (source : MétéoFrance)
	Nombreux polluants de l'eau et de l'air (non réglementés)	Particules fines à bonne résolution (grille de 100 m)	Projections climatiques futures à fine échelle (< 1 km)	Projections climatiques futures (échelle spatiale grossière, 10-100 km)
	Champs électro-magnétiques* (à confirmer)			

¹⁰⁰ Dans le cas du bruit, chaque grande agglomération est tenue, du fait de la réglementation européenne, de développer des cartographies du bruit ; ces données sont généralement sous-traitées à des bureaux d'études, distincts d'une ville à l'autre et utilisant des méthodologies non harmonisées d'une ville à l'autre et disponibles pour des années différentes. Se servir de ces données pour une étude à l'échelle du pays implique de contacter séparément chaque agglomération française, puis de faire un gros travail d'harmonisation (et implique d'exclure toutes les petites communes et zones rurales).

¹⁰¹ Dans le cas du trafic routier, il existe des données centralisées sur les grands axes nationaux, mais celles-ci ne couvrent pas les zones urbaines, pour lesquelles il faut se tourner vers chaque commune ou chaque agglomération. Ainsi, réaliser une étude sur les liens entre la proximité du trafic routier et la santé implique de se restreindre à une petite fraction du territoire et de la population, ou de réaliser un très gros travail d'harmonisation. Voir par exemple la cartographie proposée sur geoportail.gouv.fr : <https://www.geoportail.gouv.fr/donnees/plan-dexposition-au-bruit-peb>

Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé

	Fraction importante des milliers de substances chimiques sur le marché ¹⁰²			
--	---	--	--	--

Source : ITMO Santé publique

¹⁰² On ne sait pas précisément le nombre de substances chimiques présentes dans notre environnement immédiat. Le règlement européen REACH de 2006 fournit un premier niveau, rudimentaire, d'information, en listant toutes les substances commercialisées sur le marché européen par une société à plus de 1 tonne par an. Cette liste, qui ne couvre pas les médicaments, pesticides, cosmétiques ni les substances naturellement présentes dans l'environnement, a permis d'identifier environ 22 500 substances. Selon l'agence européenne de l'environnement, le nombre total de substances sur le marché dans l'UE (quel que soit le tonnage et l'usage) pourrait être de l'ordre de 100 000. Sur ces 100 000 substances, les études de biosurveillance (par exemple la surveillance alimentaire réalisée par l'ANSES dans le cadre de l'enquête EAT par exemple, ou la biosurveillance humaine réalisée par l'Anses et Santé publique France avec les études ESTEBAN puis Albane), permettent de fournir des estimations concernant l'exposition pour quelques centaines d'entre elles (en général celles pour lesquelles le niveau de preuve concernant un éventuel effet sanitaire est le plus élevé). Toutefois il faut garder à l'esprit que, dans le cas de la biosurveillance humaine, il s'agit d'enquêtes sur quelques centaines à quelques milliers de sujets, recrutées de façon à être représentatives de celles vivant en France, qui peuvent permettre de décrire le niveau d'exposition de la population ainsi que les extrêmes de cette distribution (par exemple 10ème et 90ème percentiles de l'exposition). Elles ne permettent pas des croisements avec d'autres bases de données pour tirer des conclusions sur les conséquences sanitaires de ces expositions (sauf à suivre la population des études de biosurveillance, mais dont l'effectif est trop petit pour la plupart des pathologies).

Sur certains champs, des études ad hoc restent l'approche à privilégier, faute de données suffisantes. A ce titre, les cohortes et les registres constituent des outils précieux en matière de santé environnementale.

Sur ce même champ environnemental, on peut souligner l'intérêt de la démarche du Green Data for health (présentation dans l'encadré ci-après), qui propose notamment un catalogue des bases de données existantes et vise à faciliter l'accès aux données environnementales. Le financement de quatre projets intersectoriels santé-environnement en 2023, via un appel à projet conjoint avec le Health data hub intitulé « La donnée au service de la recherche et de l'innovation en santé-environnement », est une initiative à saluer, et elle devrait prochainement être renouvelée. La mise en place d'un groupe de travail, co-animée entre le HDH et le GD4H portant sur les « données géolocalisées en santé », visant à favoriser la mobilisation et le croisement des données environnementales et sanitaires constitue également une initiative qui mériterait d'être étendue.

Le Green Data for health (GD4H)

Le « Green Data for Health » (GD4H) vise proposer une structuration des données environnementales françaises utiles pour la santé, dans un premier temps développé à l'usage des chercheurs et donc notamment de la sphère publique.

Ses objectifs, sont les suivants :

- décrire les données environnementales pertinentes et disponibles pour être croisées avec des données de santé (métadonnées, modalités d'accès, limite d'utilisation, granulométrie, contexte de collecte des données...). Un catalogue de données en ligne a ainsi été mis en ligne. Il doit encore être enrichi et sa mise à jour automatisée ;
- faciliter la « repérabilité » et l'accès aux données environnementales, y compris au niveau des territoires, dans une logique d'open data, ou, en fonction des obligations juridiques, de partage de données dans un cadre privé ou public contraint. De premiers outils juridiques ont été mis en ligne sur la qualification juridique des données ou sur le régime applicable à ces données ;
- améliorer l'interopérabilité des données environnementales entre elles et avec les données de santé et permettre un appariement des données environnementales avec les données de santé, notamment celles du système national des données de santé (« Health Data Hub »), mais aussi celles des autres acteurs de la recherche ou de l'expertise (ex : Inserm, Santé Publique France, ANSES, INERIS...). Un groupe de travail interopérabilité des données a été mis en place s'agissant des données géolocalisées en santé, coanimé entre le HDH et le GD4H.

Un deuxième enjeu réside dans le développement de l'interopérabilité des bases de données, d'autant plus complexe à réaliser à une échelle intersectorielle, comme l'illustre l'exemple des données médicales de santé au travail¹⁰³. Le développement de systèmes d'information sécurisés et interopérables entre les différents acteurs de la santé au travail habilités, essentiel pour améliorer la recherche, le suivi des salariés et assurer la continuité des parcours, s'est longtemps heurté à des blocages institutionnels et législatifs (interdiction pour les médecins du travail d'accéder aux données du dossier médical partagé - DMP, interdiction d'utiliser l'identifiant national de santé - INS, etc.).

¹⁰³ Cet exemple est tiré de travaux de l'IGAS sur la santé environnementale.

Ces freins ont été levés par la loi n° 2021-1018 du 2 août 2021 pour renforcer la prévention en santé au travail, ouvrant la voie à une mise en œuvre opérationnelle de tels SI. Outre la levée de l'interdiction d'utiliser l'identifiant national de santé, l'accès au dossier médical partagé (DMP) sera ouvert au médecin du travail qui pourra par ailleurs, sous réserve de l'accord du travailleur, l'alimenter dans le cadre d'un volet spécifique relatif à la santé au travail issu du dossier médical en santé au travail (DMST). Compte tenu des délais de mise en œuvre (notamment en termes d'implémentation des logiciels), la possibilité effective de recourir à ces données requerra toutefois encore de nombreuses années.

Le développement de structures pivots (« hubs ») sectorielles constitue le principal levier pour favoriser l'interopérabilité effective des données, en identifiant des données pivot, en favorisant l'utilisation de standards communs et en développant des thésaurus pour favoriser l'interconnexion des bases de données. Au-delà des initiatives Health data hub et Green Data for health, on peut mentionner le "Bercy hub" ou encore le projet de Labor data hub (pour les données emploi/travail), même si ces différentes structures recouvrent des réalités et des missions très variées. Une animation interministérielle de ces hubs pourrait utilement être mise en place par la Dinum, dans une optique de développement de l'interopérabilité.

Recommandation n°35 Favoriser l'utilisation de données issues de différents champs intersectoriels, notamment pour développer la prévention en santé, en s'appuyant sur les « hubs » sectoriels pour cartographier les données et favoriser leur interopérabilité

Enfin, un dernier levier d'action consiste à renforcer et faciliter **l'accès aux bases de données statistiques existantes pour les chercheurs sur l'ensemble des champs sectoriels**. Deux exemples illustrent les freins ou les difficultés à la réutilisation de données pour les chercheurs, alors même que les enjeux d'appariements sont essentiels à la recherche, par exemple pour les cohortes.

Le répertoire de gestion des carrières uniques (RGCU), qui regroupe des données sur les carrières de l'ensemble de la population française, avec des informations détaillées sur les périodes et les revenus d'activités salariés, ou d'inactivité liée au chômage et à la maladie, constitue une source de données précieuses pour des études. S'il a été conçu dans une logique d'ouverture des données, avec dès l'origine l'ambition d'une mise à disposition des données pseudonymisées du RGCU sur le CASD, son cadre réglementaire restreint actuellement sa réutilisation aux statisticiens publics, sans intégrer les chercheurs dans les publics-cibles¹⁰⁴.

De même, l'accès des chercheurs à l'EDP santé, qui est une base de données de la DREES particulièrement intéressante, puisqu'elle apparie, pour un échantillon de la population, les données du SNDS avec les données socio-fiscales de l'INSEE, reste actuellement peu lisible et complexe, avec un double circuit de demande, d'une part, auprès du comité de la statistique publique, d'autre part auprès du CESREES et de la CNIL. Selon la DREES, la source de cette complexité proviendrait du type de l'autorisation CNIL de cette base de données, l'EDP Santé ayant été autorisée et donc constituée pour évaluer la stratégie nationale de santé. Compte-tenu

¹⁰⁴Conformément au IV de l'article R 161-69-12 du code de la sécurité sociale, sont destinataires des données du RGCU, les agents exerçant des activités statistiques individuellement désignés et dûment habilités au sein des régimes et services mentionnés à l'article L. 161-17-1-2 et des services de l'Etat placés sous l'autorité respective des ministres chargés de la sécurité sociale, de l'agriculture, du budget et de la fonction publique

de la richesse de cette base de données, la mission recommande une évolution de son autorisation CNIL pour que sa réexploitation soit facilitée.

Recommandation n°36 Favoriser l'accès pour les chercheurs aux bases de données de statistique publique comme l'EDP-Santé ou le répertoire de gestion des carrières uniques (RGCU)

2.4.5.3 Renforcer l'usage des données de santé pour le pilotage des politiques publiques

En dépit de progrès significatifs, le potentiel d'utilisation des données de santé par les administrations et agences publiques pour un meilleur pilotage des politiques publiques reste encore largement sous-utilisé. Les nombreux projets de l'agence régionale de santé (ARS) d'Ile-de-France en matière d'utilisation secondaire des données de santé illustrent les apports nombreux des données de santé pour les politiques publiques.

Le renouveau de l'utilisation secondaire des données de santé par l'ARS Ile-de-France depuis le Covid

D'une manière générale, l'approche de l'Agence régionale de santé en matière d'exploitation et de réutilisation de données de santé a considérablement évolué depuis la pandémie de COVID-19. Elle est passée d'une logique plutôt centrée sur le suivi et le reporting à une approche plus intégrée :

- **Accélération du développement de la prédiction** à partir de sources de données élargies : outils spécifiques sur le COVID (anticiper l'impact sur les soins critiques) mais aujourd'hui également sur les flux aux urgences par exemple. Ces outils sont utilisés par l'Agence pour piloter l'offre, mais aussi par les établissements eux-mêmes, par les urgentistes, les structures mobiles d'urgence et de réanimation (SMUR), c'est-à-dire par une communauté plus large ;
- **Mobilisation de plus en plus systématique de l'intelligence artificielle** pour mieux comprendre un certain nombre de tendances.
- **Mobilisation des registres régionaux** (cardio, périnatalité) à des fins de travaux de recherche et d'analyse des parcours de santé (logique de « valorisation de données dormantes »). Exemple du projet Parcours Infarctus du Myocarde (projet retenu dans le cadre de l'appel à projet du Health Data Hub)
- **Passage d'une logique de coups de sonde à un dispositif de surveillance plus robuste** : les outils développés pendant la crise (SIVIC, SIDEV, Contact COVID) sont en train d'être adaptés en routine pour disposer d'outils de suivi épidémiologique beaucoup plus réactifs et et plus complets. L'effet levier sur les interventions de l'ARS Ile de France serait potentiellement considérable (adaptation de SIDEV vers Labo-SI notamment) ;
- **Elargissement du champ des données mobilisées** : exploitation plus systématique et intégrée des données en lien avec les soins non programmés, des données de périnatalité hors SNDS ou PMSI, des données de mortalité, etc.

Source : Extrait de la note de l'ARS Ile de France à l'attention de la mission

Plusieurs leviers seraient de nature à encourager une plus grande utilisation des bases de données parfois très riches (et méconnues) des administrations, comme le montrent les exemples des agences sanitaires :

- Le développement de compétences et d'expertises en matière de données est une condition *sine qua none* à leur réutilisation, et plus globalement à l'intégration des données dans les stratégies des établissements. Plusieurs agences sanitaires (Santé Publique France, ANSM, INCA) soulignent le manque de compétences pour exploiter leurs données très riches, avec un enjeu fort de recrutement de profils spécialisés et de formation des agents :
 - Le développement d'appels à projets pour l'exploitation des données est ainsi plébiscité, parce qu'il permet de financer le recrutement de compétences spécialisées, même si subsiste le défi de pérenniser les expertises à l'issue des appels à projet ;
 - Au-delà des appels à projets sur le numérique en santé dans le cadre de France 2030, on peut mentionner le programme d'Entrepreneur d'intérêt général (pilote par la direction interministérielle du numérique) qui permet l'intégration de jeunes experts du numérique (développeurs, ingénieurs de la donnée, ingénieur de la science de la donnée/data scientist...) pendant 10 mois au sein des administrations publiques pour la réalisation de projets numériques ;
- Le renforcement de l'accompagnement juridique constitue également un levier très important pour la réalisation de nombreux projets au sein des administrations. Plusieurs agences sanitaires ont souligné leur incapacité à mener des projets de réutilisation de données du fait du cadre juridique actuel, jugé très contraignant pour les croisements de données avec le SNDS, le recueil du consentement des patients ou encore l'utilisation du NIR. Par exemple, l'établissement français du sang a témoigné de sa très grande difficulté à exploiter les données des donneurs de sang du fait de l'absence d'utilisation du NIR lors des dons. L'agence de biomédecine a souligné son incapacité à exploiter son accès permanent pour des projets de recherche, rendant nécessaires les demandes d'autorisation devant la CNIL et le CESREES. Un accompagnement juridique renforcé permettrait de mieux utiliser les procédures juridiques existantes. A titre d'exemple, l'utilisation du NIR (ou de l'INS) est désormais moins contrainte qu'auparavant¹⁰⁵, mais elle doit respecter un certain nombre de règles de sécurité détaillées par la CNIL dans un guide pratique¹⁰⁶ ;

Un dialogue approfondi avec la CNIL permettrait également, le cas échéant, d'identifier les voies d'allègements possibles des procédures (élaboration de nouvelles méthodologies de référence ou adaptation des procédures existantes). La consultation prévue fin 2023-début 2024 par la CNIL pour recueillir les attentes en matière d'évolution des méthodologies de référence ou référentiels mériterait ainsi d'accorder une place particulière aux administrations et agences publiques ;

- Le développement d'outils reposant sur l'utilisation de données mérite également d'être accompagné du fait des enjeux forts en termes de pilotage des politiques (outils de

¹⁰⁵ Jusqu'en 2016, un décret en Conseil d'Etat était nécessaire.

¹⁰⁶ [Guide pratique - Modalités de circulation du NIR pour la recherche en santé aux fins d'appariement de données avec le SNDS \(cnil.fr\)](https://www.cnil.fr/fr/guide-pratique-modalites-de-circulation-du-nir-pour-la-recherche-en-sante-aux-fins-d-appariement-de-donnees-avec-le-snds)

surveillance épidémiologiques, outils de surveillance des produits de santé, outils de pilotage de l'offre de soins...). L'appui est plus particulièrement nécessaire lorsque le développement des outils suppose des appariements de données, avec le SNDS ou avec d'autres données administratives, car ces projets sont particulièrement difficiles à mettre en œuvre. A titre d'exemple, Santé Publique France ambitionne de s'appuyer sur les données des entrepôts de données de santé hospitaliers pour la surveillance épidémiologique, notamment pour les pathologies hivernales. Ce projet qui nécessite de s'appuyer sur de nombreuses sources de données, hétérogènes, avec des formats différents, est complexe mais essentiel ;

- Enfin, un accompagnement à l'ouverture des données gagnerait à être réalisé, pour favoriser la publication de données sous un format anonymisé, le cas échéant agrégé, par exemple sur le modèle de datamarts. En 2020, la mission Bothorel¹⁰⁷ soulignait que l'ouverture des données publiques constituait un levier puissant pour restaurer la confiance dans l'action publique, et permettait également d'améliorer la qualité et la fiabilité des données. La faiblesse du partage de données entre administrations publiques était aussi soulignée. Sur ce champ, on peut mentionner la publication récente du site data ansm, réalisé par l'agence avec l'accompagnement du Health data hub, qui a permis de mettre à disposition du public un grand nombre de données sur la surveillance des produits de santé, avec des outils de data visualisation (à l'instar du site data ameli de l'assurance maladie).

A ce titre, la DREES, en tant qu'administrateur ministériel des données, entend renforcer sa capacité à conseiller les administrations et agences pour utiliser la donnée à l'appui des politiques publiques. En complément, le Health data hub pourrait accompagner les administrations et agences publiques de manière opérationnelle dans leurs différentes démarches (accès, formations, conseils réglementaires, ...). Actuellement, le HDH suit une logique d'accompagnement par projets : elle accompagne ainsi 31 projets de recherche impliquant 14 administrations publiques hors établissements de santé. Compte-tenu des forts enjeux en termes de politique publique, la structuration d'un accompagnement plus systématique vis-à-vis des administrations et agences publiques doit être envisagée, au-delà de cet accompagnement par projet.

Recommandation n°37 Renforcer l'accompagnement de la DREES et du Health data hub auprès des administrations publiques et des agences sanitaires, pour favoriser l'exploitation des données de santé pour le pilotage des politiques publiques.

Par ailleurs, le pilotage des politiques publiques en cas de crise sanitaire nécessite des dispositions particulières en termes d'exploitation des données¹⁰⁸. Pendant la crise COVID, la constitution du système d'informations SI-DEP (pour la remontée des tests biologiques) a nécessité plusieurs dispositions législatives successives. La mission n'a pas approfondi spécifiquement ce sujet,

¹⁰⁷ « Pour une nouvelle ère de la politique publique de la donnée », Eric Bothorel, Stéphanie Combes, Renaud Vedel, 2020

¹⁰⁸ A ce titre, L'article 67 de la LIL prévoit que des organismes figurant sur une liste fixée par arrêté, après avis de la CNIL, peuvent mettre en œuvre des traitements dans le domaine de la santé ayant pour seule finalité de répondre, en cas de situation d'urgence, à une alerte sanitaire et d'en gérer les suites. Dans ce cas, ces traitements sont seulement soumis aux dispositions du RGPD relatives à l'analyse d'impact relative à la protection des données et à la consultation préalable de l'autorité de contrôle (CNIL).

compte-tenu des travaux actuellement menés par le Comité de Veille et Anticipation des Risques Sanitaires (COVARs), en particulier sur la question de l'interconnexion des bases de données pour la surveillance et la recherche sur les risques sanitaires. Le projet de recommandation d'un protocole permettant de prévoir « à froid » dans quelles conditions particulières l'accès aux données pourra être organisé en période de crise semble, du point de vue de la mission, une perspective toutefois très prometteuse.

Enfin, comme on l'a vu précédemment, l'utilisation des données de santé recèle un potentiel important sur le champ des produits de santé. De nombreux travaux sont actuellement menés pour favoriser l'utilisation des données de santé pour l'évaluation des produits de santé.

A titre d'exemple, l'agence de l'innovation en santé pilote un groupe de travail sur les nouvelles méthodologies d'essais cliniques (essais in silico, designs modernes comme les essais plateforme...) qui vise à évaluer la performance de ces méthodologies, à définir les cas dans lesquels il est possible d'y avoir recours et les éventuels freins à lever.

Les modèles de tarification des produits de santé vont également recourir de manière croissante aux données de vie réelle, en particulier pour les produits innovants, comme le montre le mécanisme en cours de déploiement pour les médicaments de thérapies innovantes. Ce modèle, prévu dans la loi de financement de la sécurité sociale pour 2023, repose toujours sur une négociation entre l'industriel et le comité responsable de la tarification des produits de santé, mais en y adjoignant un paramètre de paiement en fonction des résultats de vie réelle.

L'utilisation des données de vie réelle pour la réévaluation des produits de santé après leur mise sur le marché doit aussi être notablement renforcée. A ce titre, le récent rapport sur la régulation des produits de santé¹⁰⁹ préconise une réévaluation en vie réelle des produits au moins tous les 5 ans et un financement et une construction des infrastructures de données de santé le permettant (mesure E.16).

La mission rejoint aussi la proposition de construction d'un plan d'utilisation des données de santé à des fins d'évaluation en vie réelle par la DREES et le Health data hub, en associant la Haute Autorité de santé, le comité économique des produits de santé et l'agence de l'innovation en santé, ce qui permettra de prolonger, d'amplifier et de coordonner l'ensemble des travaux menés.

¹⁰⁹ « Pour un « new deal » garantissant un accès égal et durable des patients à tous les produits de santé-mission régulation des produits de santé », Agnès Audier, Claire Biot, Frédéric Collet, Anne-Aurélié Epis de Fleurian, Magali Léo et Mathilde Lignot-Leloup, 2023

LISTE DES ANNEXES

Annexe 1 : Panorama des entrepôts de données de santé

Annexe 2 : Le cadre juridique de l'accès aux données de santé à des fins de recherche

Annexe 3 : Comparaisons internationales en matière d'accès aux données de santé

Annexe 4 : Le Health Data Hub : missions et réalisations

LISTE DES PERSONNES RENCONTRÉES

Cabinets

Cabinet du président de la République

Matthieu LANDON, conseiller

Cabinet du ministre de la prévention et de la santé

Raymond LE MOIGN, directeur

Julie LAGRAVE, conseillère numérique

Cabinet du ministre de la recherche

Etienne GAYAT, conseiller spécial santé

Anne-Isabelle ETIENVRE, conseillère recherche

Clément JAKYMIW, conseiller innovation et numérique

Cabinet du ministre de l'économie, des finances et de la souveraineté industrielle et numérique

Carole VACHET, directrice adjointe du cabinet

Etienne FLORET, conseiller innovation, numérique et PME

Cabinet du ministre de l'industrie

Mathilde BOUCHARDON, conseillère santé

Directions d'administration centrale

DREES

Fabrice LENGART, directeur

Claude GISSOT, directeur de projet « administrateur ministériel des données, des algorithmes et des codes » (AMDAC)

Benoît OURLIAC, sous-directeur « Observation de la santé et de l'assurance maladie »

Direction générale de la recherche et de l'innovation

Claire GIRY, directrice générale

Isabelle ANDRE, directrice du secteur Biologie Santé (SSRI/DGRI)

Carole DUFOUIL, chargée de mission, secteur biologie et santé (SSRI/DGRI)

Délégation du numérique en santé

Raphaël BEAUFRET, co-responsable

David SAINATI, directeur innovation

Aymeric PERCHANT, directeur des projets eSanté

Direction générale de l'offre de soins

Cécile LAMBERT, cheffe de service, adjointe de la directrice générale

Marion FAGES, adjointe à la sous directrice du pilotage de la performance des acteurs de l'offre de soins

Anne LESQUELEN, cheffe du bureau produits de santé (PF 2)

Isabelle PETIT, chargée de mission au sein du bureau produits de santé

Yann-Maël Le DOUARIN, adjoint au chef de bureau « Coopérations et contractualisations » (PF3)

Albane MIRON DE L'ESPINAY, adjointe au chef de bureau innovation et recherche clinique (PF4)

Judicaël THEVENARD, chef du bureau systèmes d'information et transformation numérique (PF5)

Marie-Odile THEVENON, cheffe du bureau Données de pilotage et aide à la décision (SR5)

Direction générale de la santé

Grégory EMERY, directeur général de la santé

Marine JEAN-BAPTISTE, conseillère santé publique du directeur général de la santé

Florence LYS, adjointe de division Agences de santé, Partenariat, Concertation – DAPC, service des politiques d'appui au pilotage et de soutien

Direction générale des entreprises

Adrien LAROCHE, directeur de projet économie de la donnée, coordinateur de la stratégie cloud de France 2030

Annaëlle PARIS, chargée de mission santé numérique, chargée du suivi de la stratégie Accélération en santé numérique

Direction interministérielle du numérique

Stéphanie SCHAEER, directrice

Vincent COUDRIN, cloud transformation policy officer

Agence de l'innovation en santé

Lise ALTER, directrice

Dinh-Pong NGUYEN, responsable de la veille et de la prospective- médecin

Engueyrran HABRAN, responsable adjoint du pôle accompagnement et chargé de l'animation territoriale

Florie FILLOL, pôle accompagnement, responsable de l'accompagnement des projets innovants en santé au sein du pôle accompagnement

Benoît LABARTHE, responsable accélération recherche et du transfert en technologie

Direction de la sécurité sociale

Franck VON LENNEP, directeur

Autres opérateurs de l'État et autorité administrative indépendante

Plateforme des données de santé (Health data hub)

Gérard RAYMOND, vice-président, France asso santé

Stéphanie COMBES, directrice

Nicolas BELLIERE, chef de cabinet

Gauthier SCHREINER, directeur projets et services utilisateurs

Gil LAMP, directeur data

Louis PERRI, responsable technique de la plateforme

Pr Jean-Yves Blay, oncologue, porteur du projet DEEPSARC

Arnaud ROSIER, porteur du projet IMPLICITY

Laurence WATIER, projet BactHub, chercheuse à l'INSERM

Valérie EDEL, directrice du guichet au Health Data Hub

Caisse nationale d'assurance maladie

Thomas FATOME, directeur général

Marguerite CAZENEUVE, directrice déléguée à la gestion et à l'organisation des soins

Damien VERGE, directeur de la stratégie, des études et des statistiques (MOA du SNDS)

Caroline CHATARD-FIVAZ, responsable de la Maîtrise d'ouvrage des Systèmes d'information Sniiram-SNDS

Gildine CROIZE, responsable du département juridique

Stéphanie NAUX, responsable du pôle juridique DSES

Anne LORIN, directrice de mission à Direction Opérationnelle du Numérique et de l'Innovation en Santé.

CNIL

Marie-Laure DENIS, présidente

Louis DUTHEILLET DE LAMOTHE, secrétaire général

Hélène GUIMIOT-BRÉAUD, cheffe du service de la santé

Thomas DAUTIEU, directeur de l'accompagnement juridique

Manon de FALLOIS, adjointe à la cheffe du service de la santé de la CNIL

Erik BOUCHER, ingénieur expert – référent santé

CESREES

Bernard NORDLINGER, président

Haute Autorité de santé

Fabienne BARTOLI, directrice générale

Pierre-Alain JACHET, directeur de projet, responsable de la stratégie data

Centre d'accès sécurisé aux données

Catherine GAUDY, présidente

Kamel GADDOUCHE, directeur

GIS Epiphare

Pr Mahmoud ZUREIK, directeur

Rosemary DRAY-SPIRA, directrice adjointe

Alain WEIL, Directeur adjoint

Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail

Benoit VALLET, directeur général

Santé publique France

Caroline SEMAILLE, directrice générale

Agence de biomédecine

Marine JEANTET, directrice générale

Autorité de sûreté nucléaire

Pierre BOIS, directeur général adjoint

Agence nationale de sécurité du médicament et des produits de santé

Christelle RATIGNIER-CARBONEIL, directrice générale

BPI France

Béatrice DE KEUKELEIRE, responsable sectorielle santé, direction de l'expertise

Chahra LOUAFI, directrice du fonds Patient autonome2

ANSSI

Vincent STRUBEL, directeur général

Julie CHANTREUX, cheffe de cabinet

Silvère RUELLAN, chef du bureau santé et affaires sociales

Organismes de recherche

INSERM

Didier SAMUEL, président-directeur général

Rémi SLAMA, directeur de l'institut thématique de santé publique

Franck LETHIMONNIER, directeur de l'Institut Thématique Technologie pour la Santé

INRIA

Bruno SPORTISSE, directeur général

Cohorte Constances

Marie ZINS, responsable scientifique et technique de la cohorte, médecin épidémiologiste, enseignant-chercheur à l'Université Paris Cité, directrice de l'Unité Mixte de Service 011 « Cohortes épidémiologiques en population

Martial GOLDBERG, professeur émérite d'épidémiologie et santé publique à l'Université Paris Cité

Paris Santé Campus

Antoine TESNIÈRE, directeur général

AP-HP

Pr Raphaëlle Renard PENNA, chercheuse

Pr Olivier Steichen, professeur de médecine interne à l'Hôpital Tenon.

Pr Etienne AUDUREAU- responsable de l'URC Henri Mondor, médecin de santé publique

Autres chercheurs

Geneviève CHÊNE, porteuse de la cohorte Memento

Philippe RAVAUD, AP-HP

Ségolène AYMÉ, présidente du conseil scientifique consultatif du Health Data Hub

Parlementaires

Catherine DEROCHE, sénatrice, présidente de la commission des affaires sociales

Philippe LATOMBE, député

Associations et organismes nationaux

Haut Conseil de la santé publique

François EISINGER, président de la commission spécialisée déterminants de santé et maladies non transmissibles

CCNE

Claude KIRCHNER, directeur

Emmanuel DIDIER, sociologue, directeur de recherche au CNRS

Jérôme PERRIN, membre du comité pilote d'éthique du numérique

Conférence des présidents de CHU

Florence FAVREL-FEUILLADE, présidente de la commission Recherche et innovation

Monique SORRENTINO, DG CHU GRENOBLE

Jean-François LEFEBVRE, DG CHU TOULOUSE

Jean-Christophe CALVO, DSN CHRU NANCY

Elodie ALLARD, directrice de la recherche clinique et de l'innovation, Hospices civils de LYON

Pr Vincent VUIBLET, PU-PH CHU REIMS.

Conférence des présidents de CME

Pr Rémi SALOMON, président, président de la commission médicale d'établissement de l'AP-HP

Pr Éric STINDEL, président de la commission médicale d'établissement du CHRU de Brest

Pr Delphine MAUCORT-BOULCH, vice-présidente de la commission médicale d'établissement des hospices civils de Lyon

Comité de veille et d'anticipation des risques sanitaires

Brigitte AUTRAN, présidente

Yvanie CAILLE, membre du COVARIS

Fabrice CARRAT, membre du COVARIS

Pr Simon CAUCHEMEZ, membre du COVARIS

Dr Annabel DESGREES-DU-LOU, membre du COVARIS

Dr Roger LE GRAND, membre du COVARIS

Véronique LOYER, membre du COVARIS

Patrick GIRAUDOUX, membre du COVARIS

Pr Xavier de LAMBALLERIE, membre du COVARIS

Pr Xavier LESCURE, membre du COVARIS

Pr Rémy SLAMA, membre du COVARIS

Léa DRUET FAIVRE, chargée de mission

Mission sur la recherche clinique

Anne-Marie ARMANTERAS DE SAXCÉ

Manuel TUNON DE LARA

Charles PERSOZ, IGESR

Marc PENAUD, IGAS

Cour des comptes

Jérôme DOSSI, conseiller référendaire

Sébastien GALLÉE, conseiller référendaire

Cédric D'OLIVEIRA, conseiller référendaire

Associations de patients

Arthur DAUPHIN, chargé de mission numérique en santé de France Assos Santé

Julia CHARBONNIER, directrice d'Actions Traitements

Cédric DANIEL, chargé de mission Communications et plaidoyer chez Actions Traitements

Anne BUISSON, directrice d'Afa Crohn RCH France

Solène BAZIN, chargée de mission d'AIDES

Jean-François THEBAUT, vice-président de la Fédération Française des Diabétiques

Laurent DI MÉGLIO, administrateur de Renaloo

Manon MOLINS, responsable du projet Moi Patients chez Renaloo

Maria ROUBOTSVA, chargée de mission d'UFC Que Choisir

Lucie VIALARD, représentante de France Assos Santé au CESREES, Ligue Nationale Contre le Cancer

Serge HATCHWELL, AIDES, représentant de France Assos Santé au CESREES

Hervé NABARETTE, AFM Téléthon

Acteurs privés et leurs organismes représentatifs

Thérapixel

Pierre FILLARD, co-fondateur et CSO

SNITEM

Armelle GRACIET, directrice des Affaires Industrielles

William ROLLAND, directeur délégué au numérique en santé

Aurélie LAVOREL, responsable accès au marché et données de santé en vie réelle

Marina LEFEBVRE, juriste

Carine FAUDON-HUBNER, directrice des affaires juridiques et des questions éthiques

Brigitte CONGARD-CHASSOL, directeur des affaires médicales

François-Régis MOULINES, directeur des affaires gouvernementales

Cécile COLLIGNON, responsable des données (pharmacien) Medtronic France,

Henri SOUCHAY, responsable de la recherche clinique pour GE Health Care

Daniel SZEFTTEL, président d'une start up SEMEYA, utilisateur du SNDS

LEEM

Thomas BOREL, recherche- innovation- santé publique

Isabelle VITALI, Présidente de la Commission Numérique & Donnée de Santé (SANOFI)

Nathalie MANAUD, Directrice Innovation

Dora TALVARD, Responsable juridique données personnelles et numérique en santé

NUMEUM

Pascal BECACHE, vice-président de la commission Santé de Numeum et fondateur de Digital Pharma Lab

Mariane CIMINO, déléguée aux Usages numériques et animatrice de la commission Santé de Numeum

Charles MARIAUX, chargé de mission Usages numériques

Marine GOSSA, déléguée aux Affaires publiques

Comité stratégique de filière (Intelligence artificielle et Santé)

Corinne BLACHIER-POISSON, présidente du CSF IA et santé, AMGEN

Mélina GILBERG, Sherpa, directrice associée du pôle médico-économie et données en vie réelle, AMGEN

Virginie LASSERRE, directrice des affaires Publiques chez Janssen

Fédération nationale de la mutualité française

Yannick LUCAS, Directeur des affaires publiques

Christian PORTAFAX, Responsable du service des données de santé

Iqvia France

Jean-Marc AUBERT, président

Cardiologs

Yann FLEUREAU, co-fondateur

Owkin

Yedidia LEVY-ZAUBERMAN, responsable des affaires publiques

Agathe ARLOTTI : responsable de l'équipe partenariats

Arkhn

Cornelio MALCIU, codirigeant et cofondateur d'ARKHN

Mariuca de HILLERIN, responsable des affaires publiques

Octopize

Olivier BREILLACQ, fondateur et directeur de Octopize

Pierre-Antoine GOURRAUD, Professeur des universités et praticien hospitalier en biologie cellulaire au CHU de Nantes

Numspot/Outscale

Claire BIOT, vice-présidente de l'industrie santé, Dassault Système

Alain ISSARNI, président

Cécile LAGARDÈRE, Docaposte- pilote stratégie de santé et données

Philippe MILTIN, directeur général d'Outscale

OVH Cloud

Octave KLABA, fondateur d'OVHcloud

Anne DUBOSCQ, directrice des affaires publiques d'OVHcloud

Blandine EGGRICKX, responsable des affaires publiques d'OVHcloud

S3NS

Cyprien FALQUE, directeur général

Maxence BRICHOUX, responsable des affaires publiques

Entrepôts de données de santé

AP-HP

Nicolas REVEL, directeur général

Pr. Catherine PAUGAM-BURTZ, directrice générale adjointe

Caroline GERMAIN, directrice adjointe du pôle Innovation et données, direction des services numériques

Pierre BLONDE, directeur des services numériques par interim

Quentin DEMANET, adjoint du directeur de la recherche clinique et de l'innovation

Antoine MURZEAU, directeur BI-Pilotage, pôle Innovation et données

Thomas PETIT-JEAN, directeur (interim) de l'équipe data science, pôle innovation et données

Martin HILKA, directeur de la plateforme données massives de l'AP-HP

Dr Claire HASSEN-KHODJA, praticien hospitalier à la direction de la recherche clinique et de l'innovation de l'AP-HP

Arnaud SANDRIN, directeur opérationnel de la Banque nationale des maladies rares

CHU de Rennes

Marc CUGGIA, Professeur d'informatique médicale et PH au CHU de Rennes, directeur de l'équipe de recherche (DOMAGIA), UMR INSERM à Rennes et le centre de données cliniques du CHU de Rennes

Institut Curie

Thierry PHILIP, président

Amaury MARTIN, directeur adjoint de l'institut Curie – administrateur du GCS SeqOIA

ELSAN

Samantha PASDELOUP, directrice du développement et des partenariats, vice-présidente de l'institut des données du système de soins

Carole PUIG-CHEVRIER, directrice des affaires publiques

P4DP (médecine de ville)

Olivier SAINT-LARY, médecin généraliste à Poissy, professeur à l'université Versailles Saint-Quentin en Yvelines, président du collège national des généralistes Enseignants

David DARMON, professeur des universités à l'université Côte d'Azur, responsable du CNGE pour les données numériques

AGORiA SANTÉ

Diane-Charlotte BAILLET, directrice du consortium et collaboratrice de la société Docaposte, chef de file AGORiA

Jean-Yves ROBIN, directeur général de la société IMPACT HEALTHCARE, membre du consortium AGORiA

Personnalités qualifiées

Agnès AUDIER, Ambassadrice France 2030

Cédric Ô, ancien secrétaire d'État au numérique

Contributions écrites

Association française des CROs (clinical research organizations)
AFM-Téléthon
Agence de biomédecine
Agence nationale de sécurité du médicament et des produits de santé
Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail.
Agence régionale de santé Auvergne-Rhône-Alpes
Agence régionale de santé Guyane
Agence régionale de santé Grand Est
Agence régionale de santé Ile-de-France
Agence régionale de santé Pays-de-la-Loire
Agence régionale de santé Provence-Alpes-Côte d'Azur
Agence régionale de santé Réunion
AGORiA SANTÉ
Alliance maladies rares
Agence du numérique en santé
Agence technique de l'information sur l'hospitalisation
Collecteur analyseur de données
Conférence nationale des DG de CHU et conférence nationale des DG de CH
Conseil national de l'ordre des médecins
Conseil national de l'ordre des pharmaciens
Codoc
Dassault Systèmes
Direction générale de la santé
Docaposte
Etablissement français du sang
Euris Health Cloud
Fédération hospitalière de France
Fédération de l'hospitalisation privée
Fédération des spécialités médicales
France Assos Santé
Haute autorité de santé
Institut national du cancer
Institut de radioprotection et de sûreté nucléaire
Inserm Transfert
Leem
Pr Éric VIBERT, AP-HP
Santé publique France
Syndicat national de l'industrie des technologies médicales
Unicancer

DOCTRINE DE PRÉVENTION DES CONFLITS D'INTÉRÊTS

La mission a établi plusieurs principes pour prévenir les conflits d'intérêts au cours de ses travaux.

En premier lieu, les liens d'intérêts des membres de la mission ont été identifiés et partagés au sein de la mission. Par souci de transparence, ils sont publiés ci-dessous.

En second lieu, des règles ont été établies pour prévenir les conflits d'intérêts dans le cadre de la conduite des travaux de la mission :

- Absence de participation aux auditions d'un organisme en cas de lien d'intérêts (par exemple en cas de participation à la gouvernance de l'organisme), le membre de la mission concerné n'a pas participé à l'audition concernée ;
- Absence de prise de position sur les parties du rapport qui traiteraient / aborderaient la personne / l'organisme avec qui il y a ce lien d'intérêts.

Liste des liens d'intérêt

Stéphanie Allassonnière

Sanofi (membre d'un comité stratégique de UPCité)

Pfizer (mécène de PR[AI]RIE)

Jansen (mécène PR[AI]RIE)

Vitalaire (participation à un colloque)

Général catalysts (participation à un dîner)

Implicitity (thèse CIFRE)

Sonio (co-fondatrice)

Docaposte / Agoria Santé : membre du Comité Ethique et Scientifique

Elsan (dépôt d'un projet de création EDS)

Emilie Fauchier-Magnan

Pas de lien d'intérêt

Julien Fradel

Pas de lien d'intérêt

Anne-Sophie Jannot

Pas de lien d'intérêt

Aymeril Hoang

Membre du Comité de mission de Ramsay Santé à compter de septembre 2023

Administrateur indépendant de Cleyrop à compter du 3e trimestre 2023

Jérôme Marchand-Arvier

Pas de lien d'intérêt

SIGLES UTILISÉS

AAP	Appel à projets
AIS	Agence de l'innovation en santé
ANR	Agence nationale de la recherche
ANS	Agence du numérique en santé
ANSES	Agence nationale de sécurité sanitaire et alimentaire nationale
ANSM	Agence nationale de sécurité du médicament et des produits de santé
ANSSI	Agence nationale de la sécurité des systèmes d'information
AP-HP	Assistance publique des hôpitaux de Paris
AMI	Appel à manifestation d'intérêt
API	Application Programming Interface - interfaces de programmation applicatives
ARS	Agence régionale de santé
BPI	Banque publique d'investissement
CAD	Collecteur analyseur de données
CASD	Centre d'accès sécurisé aux données
CépiDc	Centre d'épidémiologie sur les causes médicales de décès
CESREES	Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé
CHU	Centre hospitalier universitaire
CNAM	Caisse nationale de l'assurance maladie
CNGE	Conseil national des généralistes enseignants
CNAV	Caisse nationale d'assurance vieillesse
CNIL	Commission nationale de l'informatique et des libertés
CNP	Conseil national professionnel
CPAM	Caisse primaire d'assurance maladie
CREAI	Centre régional d'Etudes d'Actions et d'Informations
CPP	Comité de protection des personnes
CSE	Comité scientifique et éthique
CSP	Code de la santé publique
CSS	Code de la sécurité sociale
COSTRAT	Comité stratégique des données de santé
COVARs	Comité de Veille et Anticipation des Risques Sanitaires
DGRI	Direction générale de la recherche et de l'innovation
DiNUM	Direction interministérielle du numérique
DM	Dispositif médical
DME	Dossier médical électronique
DMP	Dossier médical partagé
DMST	Dossier médical en santé au travail
DNS	Délégation du numérique en santé
DREES	Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
DPO	Délégué à la protection des données
EDS	Entrepôts de données de santé
ENS	Espace numérique de santé
EHDS	European Health Data Space (<i>Espace européen des données de santé</i>)
ESND	Echantillon du système national des données de santé

FDA	<i>Food drug administration</i>
FIAC	Filière d'intelligence artificielle et cancers
GD4H	Green Data for Health
GCS	Groupement de coopération sanitaire
HAS	Haute autorité de santé
HDAB	Organisme responsable de l'accès aux données de santé
HDH	Health data hub (plateforme des données de santé)
IA	Intelligence artificielle
IGAS	Inspection générale des affaires sociales
IGESR	Inspection générale de l'Éducation, du Sport et de la Recherche
INCA	Institut national du cancer
INRIA	Institut national de recherche en sciences et technologies du numérique
INS	Identifiant national de santé
INSEE	Institut national de la statistique et des études économiques
INSERM	Institut national de la santé et de la recherche médicale
IRES	Institut pour la recherche en santé publique
MDPH	Maison départementale des personnes handicapées
MR	Méthodologie de référence
NIR	Numéro d'inscription au répertoire
OCDE	Organisation de coopération et de développement économiques
ONDAM	Objectif national de dépenses d'assurance maladie
ORS	Observatoire régional de santé
OTT	Offices de transfert de technologies
PME	Petite et moyenne entreprise
PMSI	Programme de médicalisation des systèmes d'information
REIN	Réseau Epidémiologie et Information en Néphrologie
RGCU	Répertoire de gestion des carrières uniques
RGPD	Règlement européen sur la protection des données personnelles
SASN	Stratégie Santé numérique
SATT	Société d'accélération du transfert de technologies
SGMAS	Secrétariat général des ministères chargés des affaires sociales
SIGAPS	Système d'interrogation, de gestion et d'analyse des publications scientifiques
SMUR	Structure mobile d'urgence et de réanimation
SNDS	Système national des données de santé
SNIIRAM	Système national d'information inter-régimes de l'assurance maladie

LETTRE DE MISSION



GOVERNEMENT

*Liberté
Égalité
Fraternité*

Paris, le **13 1 MAI 2023**

Le Ministre de l'Economie, des Finances et de la
Souveraineté industrielle et numérique

La Ministre de l'Enseignement Supérieur et de la
Recherche

Le Ministre de la Santé et de la Prévention

A

Monsieur Jérôme MARCHAND-ARVIER,
Conseiller d'Etat

Objet : Lettre de mission relative aux données de santé

L'article 41 de la loi Organisation et Transformation du Système de Santé, promulguée le 24 juillet 2019, a élargi le Système National des Données de Santé (SNDS) à toutes les données associées à un remboursement de l'Assurance Maladie. Ce faisant, le législateur a entendu permettre un accès plus large, bien que réglementé, à un patrimoine national de données d'une importance majeure pour la recherche en santé et l'innovation.

Les changements voulus portent sur de nombreux aspects, parmi lesquels le périmètre, la gouvernance ou l'organisation de la mise à disposition. Ils ont nécessité la déclinaison d'un ensemble de textes réglementaires aujourd'hui effectifs, même si leur élaboration et leur validation a été retardée par la crise Covid-19.

Le dispositif du SNDS élargi est désormais composé, d'une part, de la base principale du SNDS¹ et, d'autre part, d'un « catalogue de données ». Ce catalogue, collection de bases issues de diverses cohortes et plus généralement d'entrepôts de données plus médicalisées ou plus spécifiques à une pathologie que la base principale, est construit de manière progressive et itérative. Les bases de données du catalogue devront pouvoir être appariées au niveau individuel avec la base principale pour un potentiel d'exploitation décuplé.

La loi de 2019 a en outre créé un comité stratégique des données de santé, ayant pour mission d'élaborer les orientations stratégiques de la politique publique en matière d'utilisation des données de santé à des fins de recherche et d'innovation (prévention, progrès thérapeutiques, qualité des soins) ainsi que d'aide au pilotage et à la régulation de l'offre de soins. Ce comité stratégique, présidé par le ministre en charge de la santé, est aujourd'hui co-animé par le ministère en charge de la recherche. Pour exploiter au mieux le potentiel du SNDS, les pouvoirs publics ont également inscrit dans la loi la création de la Plateforme des données de santé, dite « Health Data Hub » (HDH). Groupement d'intérêt public regroupant l'ensemble des acteurs publics et privés, le HDH a pour objectif de faciliter et centraliser le partage des données de santé issues de sources très variées, afin de favoriser la recherche et l'innovation. Son action s'articule autour de quatre activités principales : un guichet administratif pour accompagner les porteurs de projets dans leurs démarches, la mise à disposition d'une plateforme sécurisée à l'état de l'art, offrant des capacités avancées de stockage, de calcul, de rapprochement et d'analyse des données, l'accès aux bases de données du « catalogue de données », l'animation enfin d'une démarche autour du partage des connaissances et des outils favorisant l'utilisation secondaire des données.

¹ (SNIIRAM/PMSI/Causes de décès, données des MDPH en passe d'être rendues disponibles courant 2023)

En un peu moins de quatre ans, d'importants jalons ont été franchis : outre la production des différents textes réglementaires, la première version du catalogue avec une dizaine de bases de données référencées a été publiée. La plateforme des données de santé a mis en place et développé l'animation autour de l'utilisation des données, et renforcé le dialogue avec les citoyens via une équipe dédiée. Elle accompagne 80 projets et a conclu une centaine de partenariats actifs. Elle est non seulement reconnue par l'écosystème français, mais aussi par les acteurs européens et internationaux (Commission européenne, Harvard, MIT, ...).

De nombreux travaux restent néanmoins à conduire pour poursuivre et intensifier l'usage des données, renforcer la coordination des acteurs dans le domaine et libérer le potentiel d'innovation du secteur. Beaucoup d'entreprises doivent par exemple effectuer leur recherche et développement sur des données étrangères (ex : Cardiologs, Owkin, etc.) et privilégient un développement à l'international plutôt qu'en France, pouvant ainsi retarder l'arrivée d'innovation importante pour les patients français.

Le comité stratégique a déjà lancé plusieurs groupes de travail, chargés d'élaborer briques par briques le cadre qui permettra de construire dans la durée le patrimoine partagé des données de santé et de faciliter ses usages : estimations des coûts d'investissement et d'exploitation des entrepôts hospitaliers de données de santé recueillies dans le cadre des soins, modalités de leur financement, socle minimal de données à recueillir dans les différents entrepôts de données de santé et harmonisation des modalités d'accès aux données.

Une meilleure utilisation secondaire des données de santé est essentielle pour accélérer la recherche et l'innovation en santé en France, notamment dans le développement et la validation d'outils d'intelligence artificielle, dans l'évaluation de l'efficacité en vie réelle des produits et interventions en santé, ainsi que pour le pilotage et l'amélioration continue des politiques publiques et du système de santé. C'est un enjeu de premier plan pour la santé des personnes.

Cette montée en puissance constitue également un enjeu majeur pour l'économie française en matière d'innovation, de compétitivité et d'attractivité. Le Plan Innovation Santé 2030, notamment via la Stratégie d'accélération « Santé numérique », apporte dans cette perspective des moyens d'action inédits aux acteurs de l'innovation numérique en santé, chercheurs et industriels de toutes tailles, qui expriment des attentes fortes en matière de mise à disposition rapide des données de santé.

Compte tenu de ces enjeux, nous souhaitons que vous pilotiez les travaux ci-dessous pour le compte et en complément de ceux déjà menés par le comité stratégique des données de santé pour :

- Catégoriser et chiffrer les opportunités associées à l'utilisation secondaire des données, à la fois en termes de santé publique, d'innovation et de compétitivité et identifier sur cette base les domaines prioritaires de données qu'il serait utile de collecter et/ou de croiser. Il s'agira notamment de répondre aux futurs enjeux en termes de santé publique en regard des priorités de la stratégie nationale de santé, du basculement vers une santé plus préventive, plus personnalisée, ainsi qu'en termes d'innovation et de compétitivité. L'initiative prise par le ministère de la santé et de la prévention pour développer les entrepôts de données hospitaliers a permis de lancer une dynamique de constitution de données à partir des dossiers patients informatisés des établissements. Il est sans doute souhaitable d'ouvrir d'autres champs de recueil, mais aussi de les croiser avec des données en provenance d'autres domaines (environnement notamment) pour contribuer à la prise en charge de la santé de manière globale (approche « one health ») ;
- Cartographier la dynamique actuellement à l'œuvre de constitution de grands entrepôts de données de santé. Des initiatives multiples, isolées ou sous forme d'alliances, cherchent à capitaliser sur les données de santé en constituant des entrepôts ou des plateformes de mise à disposition (France Cohortes, OncoDataHub, Paris-Saclay Cancer Cluster, Agoria...). Tous ces acteurs sollicitent sous une forme ou sous une autre des financements publics. Il est nécessaire de collecter et de cartographier l'ensemble des initiatives en cours, d'examiner la façon dont elles s'articulent, pour dégager du côté de l'action publique les actions prioritaires à sécuriser et à engager dans les tous prochains mois ;

- Identifier les impacts organisationnels pour les différents acteurs du futur règlement sur l'espace européen de données de santé (EHDS), redéfinissant les modalités de l'usage secondaire des données de santé, entre détenteurs et utilisateurs des données, au travers d'organismes responsables de l'accès aux données de santé (HDAB), et créant une infrastructure européenne, pour laquelle le HDH a pris la tête du consortium pilote ;
- Apporter une analyse prospective sur les conditions d'un développement ambitieux de l'innovation autour des données de santé (recueil, exploitation, diffusion -, mobilisation de ces données via data science/IA, modalités de soutien des startups) notamment dans les priorités préalablement identifiées afin de faire de la France un leader international dans ce domaine, en soulignant les enjeux de modèle économique sur les différentes phases du cycle de la donnée. Sur ce dernier point, vous pourrez vous appuyer notamment sur les travaux du GT financement du comité stratégique des données de santé, afin d'explicitier les caractéristiques d'un modèle économique vertueux, favorisant les échanges de données et la structuration d'un maillage d'EDS de qualité. Devront notamment être étudiées les conditions de rémunération des acteurs mettant à disposition leurs données et les modalités d'un soutien financier public permettant d'atteindre les objectifs poursuivis.

A cet effet, vous animerez une mission composée de Mme Anne-Sophie Jannot, maître de conférences des universités-praticien hospitalier à l'APHP, Mme Stéphanie Allassonnière, professeur et vice-présidente valorisation à l'université Paris Cité et de M. Aymeril Hoang, indépendant. Vous procéderez à l'ensemble des auditions qui vous paraîtront utiles. Vous associerez également les membres du comité stratégique et vous pourrez compter sur l'appui des services de la DREES, de la DGOS, de la DNS, de la DGRI, de la DGE, de l'INRIA et du HDH. Nous souhaiterions pouvoir disposer des résultats de vos travaux d'ici la fin du mois de septembre 2023 afin que vos propositions jettent les bases d'une véritable feuille de route française en matière d'utilisation secondaire des données de santé, sur laquelle le comité stratégique des données santé a engagé des réflexions et qu'il adoptera pour les cinq années à venir.

Sachant pouvoir compter sur votre mobilisation, nous vous prions d'agréer l'expression de nos sincères considérations.



Bruno LE MAIRE



Sylvie RETAILLEAU



François BRAUN