Avis du Comité de Veille et d'Anticipation des risques sanitaires (COVARS)

du 26 juin 2024

Sur le développement, la gouvernance et l'accès aux bases de données de santé humaine en anticipation des crises sanitaires

Membres du Comité de Veille et d'Anticipation des Risques Sanitaires associés à cet avis :

Brigitte AUTRAN, Présidente, Immunologiste
Fabrice CARRAT, Épidémiologiste
Yvanie CAILLE, Association de patients
Simon CAUCHEMEZ, Modélisateur
Julie CONTENTI, Urgentiste
Annabel DESGREES du LOU, Démographe
Didier FONTENILLE, Entomologiste
Patrick GIRAUDOUX, Eco-épidémiologiste,
Mélanie HEARD, Politiste en santé
Xavier de LAMBALLERIE, Virologue
Thierry LEFRANCOIS, Vétérinaire,
Roger LE GRAND, Vaccins,
Xavier LESCURE, Infectiologue
Bruno LINA, Virologue
Véronique LOYER, Représentante des citoyens
Denis MALVY, Infectiologue
Céline OFFERLE, Association de patients
Olivier SAINT-LARY, Généraliste
Rémy SLAMA, Épidémiologiste

Cette note a été co-pilotée par Fabrice CARRAT, Simon CAUCHEMEZ et Rémy SLAMA, avec le soutien rédactionnel de Léa Druet-Faivre.

COVARS
Comité de veille et d'anticipation des risques sanitaires

Cet avis a été transmis aux autorités nationales le 26 juin 2024

Comme les autres avis du Comité de Veille et d'Anticipation des Risques Sanitaires, cet avis a vocation à être rendu public.

Auditions conduites par le COVARS

Le présent Avis a été rédigé grâce à l'appui d'une série d'auditions des acteurs et professionnels suivants, que le COVARS remercie.

-Le 26 juin 2023, M. Fabrice LENGLART, Directeur de la **Direction de la recherche, des études, de l'évaluation et des statistiques (DREES)** – Service statistique ministériel dans les domaines de la santé et du social, et Mme Mathilde GAINI, Sous-directrice du suivi et de l'évaluation des politiques de l'emploi et de la formation professionnelle à la **DARES**.

-Le 26 juin 2023, la **CNIL,** Direction de l'accompagnement juridique, en présence de Mme Valérie PEUGEOT, Commissaire en charge des données de la santé ; Mme Manon de FALLOIS, Adjointe à la cheffe du service de la santé ; et M Erik BOUCHER de CREVECOEUR, Ingénieur référent santé au service de l'expertise

-Le 29 juin, M. Damien Vergé, Directeur de la stratégie, des études et des statistiques de la **Caisse Nationale de l'Assurance Maladie**

-Le 3 juillet, **Santé Publique France**, en présence de Mme Laetitia HUIART, Directrice scientifique et M. Yann LESTRAT, Directeur de la Direction Data

-Le 3 juillet, **ANSM – GIS EPI-PHARE**, en présence de Pr Mahmoud ZUREIK, Directeur GIS EPI-PHARE et Dr Rosemary DRAY-SPIRA, Directrice adjointe GIS EPI-PHARE, Directrice de recherche Inserm

- Le 9 octobre 2023, la **DGS**, en présence du Dr Gregory Emery, Directeur Général de la Santé, Marie Baville, sous-directrice de la veille de et de la sécurité sanitaire, et Clément Lazarus, conseiller-expert auprès du DGS

-Le 23 octobre 2023, Jérôme Marchand-Arvier, Conseiller d'État en charge de la **mission *Données de santé*** qui lui a été confiée par M. le Ministre François Braun le 17 mai 2023, ainsi qu'Anne-Sophie Jannot, Émilie Fauchier-Magnan et Stéphanie Allassonnière

-Le 6 novembre 2023, **la Délégation au Numérique en Santé (DNS),** représentée par Xavier Vitry, responsable d'un pôle portant sur les SI mobilisables par temps de crise

-Le 20 novembre 2023, Stéphanie Combes, directrice du **Health Data Hub**

-Le 4 décembre 2023, les chercheurs Pr Odile Launay et Dr Liem Binh Luong Nguyen (Respivac, Cov-popart)

-Le 21 décembre 2023, la **UK Health Security Agency** en présence de Steven Riley, Director General Data, Analytics and Surveillance

-Le 21 décembre 2023, **Statens Serum Institute Denmark – Division of Infectious Disease Preparedness**, en presence de Tyra Grove Krause, Specialist in Public Health Medicine et de Marianne Voldstedlund, head of department for digital integration and analysis

-Le 8 février 2024, **France Asso-Santé** (Gérard Raymond, président, et Arthur Dauphin, chargé de mission "numérique en santé") et **AIDES** (Catherine Aumond, secrétaire générale et Solenn Bazin, chargée de mission Accès Santé)

-Le 15 mars 2024, **ANSES**, son Directeur Général le Pr. Benoît Vallet, Matthieu Schüler, Jean-Luc Volatier et Lucie Moreels.

Table des matières

Avant-Propos

Plusieurs rapports récents ont abordé la question des données de santé :

- Un rapport de juillet 2023 du Sénat sur les données de santé (rapporteuse, Mme Deroche[1]),
- Un rapport de décembre 2023 piloté par M. Marchand-Arvier à la demande des Ministres en charge de l'économie et des finances, de l'enseignement supérieur et de la recherche, et de la santé et la prévention, sur l'utilisation secondaire des données de santé[2].

En comparaison des rapports déjà publiés, les spécificités du présent Avis du COVARS sont notamment :

- De partir de la problématique des données mobilisables en temps de crise sanitaire et de l'analyse de la crise Covid-19 ;
- De ne pas se focaliser sur la problématique de l'utilisation secondaire des données de santé, qui nous semble trop restreinte, voire constituer un biais de cadrage ;
- La vision du développement, de la gouvernance et de l'accès aux données de santé en anticipation des crises sanitaires adoptée dans l'Avis du COVARS s'appuie sur la démarche scientifique classique allant 1) de la définition d'un problème ou d'une grande question de santé publique, 2) à l'identification des outils et données permettant éventuellement d'y répondre, 3) à la mise en œuvre, si besoin, d'un système ou d'une étude permettant de recueillir les données adéquates (ou demander accès aux données pertinentes existant éventuellement), 4) à l'analyse de ces données, et enfin 5) à la synthèse de toutes les connaissances, provenant de différentes disciplines, sur le sujet pour aider à la décision publique. Cette vision est confortée par notre analyse de la crise Covid-19. Elle conduit à remettre en cause le principe plus ou moins explicite du pilotage du champ des « données de santé » sous l'angle principalement technique, par des acteurs spécialisés dans les systèmes d'information et sans expérience forte en santé et santé publique, reléguant au second plan les grands acteurs de la prévention, santé publique, surveillance et recherche ;
- De ne pas traiter indistinctement tous les types de données de santé mais, en cohérence avec le point précédent, de reconnaître les spécificités des quatre grands champs que sont les données du soin (correspondant aux données de santé au sens strict de la loi), les données de surveillance en santé, les données de la statistique publique (déjà très bien structurées mais assez difficilement accessibles) et les données de recherche en santé (moins bien structurées).

Par ailleurs, dans son rapport sur l'expertise publique en situation de crise de 2022[3], la Haute Autorité de Santé a produit des recommandations spécifiques liées aux systèmes d'information. Notre rapport partage l'essentiel des constats – notamment ceux issus de l'analyse de la crise Covid-19 – et reprend, développe ou complète l'essentiel des recommandations produites par la HAS.

Le COVARS insiste sur la nécessité de changer de paradigme en plaçant les producteurs et utilisateurs de données au centre de la gouvernance et en augmentant la place des acteurs de la recherche dans celle-ci.

[1] https://www.senat.fr/notice-rapport/2022/r22-873-notice.html
[2] Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé
https://sante.gouv.fr/IMG/pdf/rapport_donnees_de_sante.pdf
[3] https://www.has-sante.fr/upload/docs/application/pdf/2023-02/lexpertise_publique_en_sante_en_situation_de_crise_-_rapport_danalyse_prospective_2022.pdf

Résumé exécutif

Contexte et objectifs : Comme illustré lors de la crise Covid-19, les données de santé sont essentielles pour suivre en temps réel l'évolution d'une situation épidémique, identifier les facteurs de risque, évaluer l'impact des interventions et traitements et modéliser l'impact potentiel de nouvelles mesures de gestion, permettant de guider la décision publique. L'objectif de cet avis du COVARS est de tirer les leçons des crises sanitaires récentes concernant l'apport de ces données et les limites rencontrées dans leur utilisation et de faire des recommandations pour améliorer le recueil et l'utilisation de ces données lors de crises futures, mais également hors du contexte de crise.

Constat *:* Après auditions des acteurs concernés, le COVARS fait les constats suivants : i) le terme de données de santé est utilisé avec des sens différents selon les acteurs ; ii) les champs contribuant aux données sur la santé (surveillance, statistique publique, soin, recherche) sont inégalement structurés ; iii) la gouvernance actuelle des données de santé est centrée sur les aspects techniques, certes importants, mais qui conduisent à reléguer au second plan les objectifs associés à ces données que sont les questions de santé publique, de recherche et d'innovation ; iv) les principaux acteurs contribuant à la production et à l'analyse des données de santé sont peu coordonnés et leurs rôles pas toujours explicites ; v) le recueil et l'interconnexion de données de surveillance et du soin utiles pour guider la décision publique, comme les causes d'hospitalisation en réanimation, ou le lien entre génome viral et données cliniques – ne sont pas organisés ; vi) durant la pandémie de la COVID-19, la mise en place de grandes études de recherche, qui se sont révélées informatives, a été lourde, difficile et tardive pour partie en raison de l'absence d'infrastructures préexistantes ; vii) la surveillance environnementale est peu développée et structurée.

Recommandations : Le COVARS recommande notamment de :

1- Mieux structurer la gouvernance des données de santé et l'activité des différents acteurs contribuant aux connaissances utiles pour la santé et la gestion de crise sanitaire, en favorisant le dialogue entre surveillance, statistique publique, soin et recherche, en remettant les utilisateurs et producteurs de connaissances (agences sanitaires, chercheurs, société civile et associations de patients) au cœur du dispositif des données de santé et **en le structurant autour de grandes questions et finalités de recherche, santé publique et innovation, plutôt que sous le concept technique de données de santé** ;

2- Enrichir les données de santé et renforcer les capacités d'appariement :

- **Enrichir la base principale du SNDS** en données cliniques, biologiques, sociodémographiques et sur de grands facteurs de risque (tabac, corpulence) et améliorer sa fréquence d'actualisation,
- **Développer la capacité à apparier et enrichir les bases de données**, notamment en systématisant l'utilisation d'un identifiant unique dans les bases nationales, les données issues de la recherche et toute autre base susceptible d'être appariée ;

3- Faciliter et simplifier l'accès aux données à des fins de surveillance épidémiologique, d'évaluation, de recherche et d'aide à la décision publique, et alléger les procédures administratives et réglementaires, principales causes d'allongement des délais d'accès ;

4- **Anticiper dès à présent, en inter-crise, le système d'information à déployer lors de futures situations de crise sanitaire ;**

5- **Mieux impliquer les citoyens et les patients,** via les instances de la démocratie sanitaire, dans l'ensemble du système des données de santé, mais aussi en amont dès la mise en place des études et bases de données, dans la gouvernance, les études et la valorisation des résultats ;

6- **Innover dans la génération de données de santé en anticipation des crises sanitaires et développer la recherche sur la surveillance de la santé et l'environnement** et, plus généralement, soutenir fortement la recherche en biologie-santé-environnement afin notamment d'élargir le vivier des experts capables de générer et analyser des données sur la santé pertinentes en temps de crise sanitaire.

Acronymes

ANSM : Agence nationale de sécurité du médicament et des produits de santéANSES: Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail

AIPD : Analyse d'impact relative à la protection des données

CEPIDC : 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

CNIL : Commission nationale informatique et libertés

CNAV : Caisse Nationale d'Assurance Vieillesse

CNR : Centres nationaux de référence de l'Institut Pasteur

CSNS : Code Statistique non signifiant

DGOS : Direction générale de l'Offre de Soins (direction du Ministère de la santé)

DGS : Direction générale de la santé (direction du Ministère de la santé)

DNS : Délégation au Numérique en Santé

DGRI : Direction générale de la recherche et de l'innovation (direction du Ministère de la recherche)

DREES : Direction de la recherche, des études, de l'évaluation et de la statistique (direction du Ministère de la santé)

EDS : Entrepôt de données de santé

EFS : Établissement Français du Sang

EHDS : European Health Data Space

HAS : Haute Autorité de Santé

HCSP : Haut Conseil de Santé Publique

HDH : Plateforme des données de santé (ou PDS)

INSEE : Institut national des statistiques et des études économiques

Inserm : Institut national de la santé et de la recherche médicale

MDO : Maladie à déclaration obligatoire

MESR : Ministère de l'enseignement supérieur et de la recherche

MR : Méthodologie de référence

NIR : Numéro d'Inscription au Répertoire national d'identification des personnes physiques

OSCOUR : Organisation de la surveillance coordonnée des urgences

PDS : Plate-forme des données de santé (ou HDH)

PMSI : Programme de médicalisation des systèmes d'information

RGPD : Règlement général sur la protection des données

SI : Systèmes d'information

SNDS : Système national des données de santé

SNGI : Système National de gestion des identifiants

SpF : Santé publique France (Agence nationale de santé publique)

SURSAUD : Surveillance syndromique des urgences et des décès

WNV : Virus du West Nile

I. Les recommandations du Covars

Le principe guidant l'élaboration de ces recommandations sur la production et l'utilisation des données de santé durant les crises sanitaires futures est de réussir à développer, documenter, interconnecter et faciliter l'ouverture des systèmes d'informations et bases données de santé humaine hors période de crise. Cela doit permettre de limiter la création de systèmes d'information *ad-hoc* lors de chaque situation d'émergence ou crise sanitaire, laquelle doit être anticipée et simple pour pouvoir être efficace.

1. **Améliorer et structurer la gouvernance des données de santé en y renforçant la place des producteurs et utilisateurs**

1.1 Revoir l'articulation et les rôles des différents acteurs dans la production et l'exploitation des données de santé, dont le nombre et la redondance possibles nuisent à l'efficacité du système.

1.2 Mettre au premier plan les producteurs et utilisateurs de données de santé (agences sanitaires, en particulier SpF, chercheurs, citoyens et associations de patients), plutôt que des institutions plus techniques centrées sur le numérique, afin qu'ils soient au cœur **d'une gouvernance stratégique des données de santé** devant **:**

- **fonctionner tant en temps de crise qu'entre les crises,**
- **identifier les priorités par grandes familles d'objectifs, grands types de données manquants, et ceci en synergie avec les homologues de l'UE, notamment dans la perspective de l'Espace Européen des Données de Santé.**
- **réunir les acteurs** des cinq domaines que sont : i) la **surveillance** (autour de Santé publique France (SpF), de l'ANSM), ii) les **données du soin** (autour de l'Assurance Maladie et/ou du HDH), iii) la **statistique publique** (autour de la DREES), iv) **la recherche en santé** (autour du MESR et de l'Inserm), et v) la **surveillance de l'environnement** (autour de l'ANSES). En situation de crise liée à un agent infectieux, SpF pourrait être l'acteur central de cette gouvernance.

1.3 S'assurer que chacun de ces acteurs bénéficie des moyens opérationnels et techniques nécessaires, ce qui n'est pas le cas à l'heure actuelle notamment pour SpF et les acteurs de la recherche. Des acteurs comme le HDH ou la DNS doivent pouvoir assurer un soutien technique.

1.4 Mieux organiser et structurer le monde de la recherche en santé concernant les données qu'il génère et renforcer le dialogue et les collaborations avec les autres grands domaines des données de santé (surveillance, soin, statistique publique, surveillance de l'environnement). **C**ette structuration doit permettre une meilleure priorisation et identification des besoins en matière de données, la mise en place d'études synergiques et communicantes pour éviter les initiatives redondantes, la définition d'une politique de partage des données, un dialogue renforcé entre tous les acteurs de la recherche en santé, une amélioration de la documentation et de la qualité des données produites.

1.5 Faire converger et dialoguer davantage la politique de *science ouverte* promue par le Ministère en charge de la recherche **et la politique concernant les données de santé,** surtout pilotée au Ministère en charge de la santé.

2. Enrichir les données de santé et renforcer les capacités d'appariement

Enrichissement :

2.1 Réaliser une cartographie des données de santé existantes et un schéma directeur des données à recueillir et apparier en vue d'enrichir le SNDS. Ce schéma directeur devrait en particulier enrichir le SNDS avec i) les résultats des analyses biologiques via l'entrepôt national des données de biologie - laboé-SI, ii) les dossiers cliniques des entrepôts de données de santé (EDS), iii) le dossier pharmaceutique, d) les données de cabinet et d'imagerie de ville, iv) les données génomiques, f) les données issues de la télémédecine, v) des données socio-démographiques individuelles et sur les grands facteurs de risque (tabac, alcool, corpulence…), vi) les certificats de naissance, certificats du 8ème jour, les examens obligatoires à 3, 6 et 9 mois, vii) les données de vaccination, quel que soit le lieu de la vaccination (école, EHPAD, travail…).

2.2 Renforcer et structurer le réseau d'entrepôts de données de santé hospitaliers, qui représentent un atout majeur en matière de données cliniques, mais dont la constitution est marquée par une vaste hétérogénéité et des conditions d'accès très contraignantes. **Créer un système d'information national** permettant de suivre les hospitalisations dans les services de réanimation des établissements de santé et un second dans les établissements sociaux et médico-sociaux, ce qui n'est actuellement pas possible.

2.3 Améliorer la qualité de la donnée par une systématisation des processus et standardiser les référentiels utilisés à travers tous les systèmes d'information de données de santé (notamment référentiel géographique, terminologie des maladies…) afin de faciliter leur (ré)-utilisation au sein du SNDS enrichi.

2.4 Développer fortement les données relatives à l'environnement au sens large (expositions professionnelles, santé animale et végétale, contamination de l'eau et de l'air, de l'alimentation, cadastre d'usage des pesticides, composition des produits de consommation courante...) et favoriser leur dialogue et interopérabilité avec les données de santé.

2.5 Soutenir la réalisation de ces objectifs d'enrichissement par la **mise en place de moyens humains et matériels dimensionnés.**

Appariement :

2.6 Rendre systématique l'utilisation d'un identifiant unique dans les différents systèmes d'information qui permette l'interconnexion de plusieurs bases de données comme le font des voisins européens notamment le Danemark. Cet identifiant devrait être le NIR, option rapide, sûre et efficace.

2.7 Améliorer la qualité du NIR en agissant 1) ***en amont*** sur la collecte primaire afin de minimiser le risque d'erreur (par exemple, éviter les saisies manuelles multiples) et 2) ***en aval :*** développer et diffuser des algorithmes pour valider le NIR.

***2.8* Réduire les délais d'appariement** avec la base principale du SNDS qui, par sa quasi-exhaustivité et son chaînage entre données de consommation de soins (DCIR), d'hospitalisation (PMSI) et des décès, est le pilier central d'un système interconnecté. L'appariement avec les EDS hospitaliers, les données des laboratoires d'analyse, les données d'accès précoce, les données de cohorte/recherche clinique, registres, systèmes de surveillance et bases publiques spécifiques (exemple de l'EDP-santé à la Drees) doit être rendu opérationnel et faciliter l'accès à ces ressources.

Fréquence d'actualisation :

2.9 Mettre en place un dispositif de remontée en flux tendu des données du PMSI en situation d'émergence, à l'instar de ce qui est développé via le PMSI *fast-track*. Améliorer l'accès aux données du PMSI *fast-track,* qui n'est pas accessible pour tous les projets de recherche. Améliorer également les délais de réception du PMSI *fast-track*, qui est délivré avec plusieurs mois de décalage.

2.10 Progresser sur la remontée du statut vital, qui devrait être disponible très rapidement dans le SNDS, et continuer à progresser sur les causes médicales de décès.

3. Faciliter et simplifier l'accès aux données à des fins de surveillance épidémiologique, d'évaluation, de recherche et d'aide à la décision publique, et alléger les procédures administratives et réglementaires

Des évolutions de fond sont indispensables afin de renforcer les utilisations secondaires des données de santé et de permettre la réutilisation des données produites dans le cadre des soins à des fins de surveillance épidémiologique, d'évaluation, de recherche et d'aide à la décision publique. En ce qui concerne l'accès aux données de santé sur projet de recherche, **une réduction des délais** techniques, réglementaires et administratifs d'accès et d'appariement aux données doit être mise en œuvre.

3.1 Clarifier les obligations relatives à l'information des patients ou des citoyens, la multiplicité des démarches en fonction des cas de figures rendant peu aisée la compréhension des exigences réglementaires pour les utilisateurs. En particulier, organiser et simplifier l'obligation d'information individuelle des patients ou citoyens en cas de réutilisation secondaire de données. Cela peut passer par la mise en place d'un portail de transparence unique connu de tous et facilement accessible.

3.2 Mettre en place des procédures plus efficaces et transparentes pour les utilisateurs qui accèdent au SNDS, via les organismes publics disposant d'un accès permanent à la base principale dans le cadre de leur mission d'intérêt public (par exemple par l'application d'une charte de bonne pratique se substituant aux procédures internes dans les organismes utilisateurs).

3.3 Simplifier les modalités d'application des obligations règlementaires prévues dans le RGPD de sorte que celles-ci soient à la fois efficaces et rapidement mise en œuvre ; renforcer les moyens de la CNIL, des organismes concernés et du CESREES ; développer une méthodologie de référence (MR) « cohorte ».

3.4 Limiter au strict minimum les contrats entre partenaires publics impliqués sur un même projet. Développer des procédures standardisées et simplifiées pour la négociation, la rédaction et la signature des conventions entre les administrations. Cela pourrait inclure la création de modèles de conventions pré-approuvés pour des cas courants de réutilisation des données de santé.

3.5 Établir des délais de traitement prédéfinis pour chaque étape du processus de mise en place de conventions juridiques entre les administrations, en veillant à ce qu'ils soient réalistes et respectés.

3.6 Renforcer les expertises techniques et juridiques nécessaires pour constituer, entretenir et apparier les bases de données utiles au progrès de la connaissance chez tous les acteurs de la recherche et de la surveillance. Cela implique notamment de revoir les grilles de salaire, en particulier chez les agents titulaires de la fonction publique, étant donné la tension existant sur ces métiers.

3.7 Étoffer le dispositif de formations (initiale/continue) sur tous ces métiers.

4. Anticiper dès à présent le système d'information à déployer lors de futures situations de crise sanitaire

La préparation de la réponse aux crises sanitaires se construit en amont des crises, en faisant évoluer les systèmes de surveillance existants et en augmentant leur agilité. Il ne s'agit pas de construire des systèmes uniquement pour le cadre de l'émergence, mais au contraire de construire une surveillance qui soit adaptable et puisse être réutilisée et « mise à l'échelle », avec un fonctionnement modifié lors d'une pandémie (par exemple, en accroissant la fréquence d'actualisation).

4.1 Engager une réflexion sur les modifications des systèmes des données de santé en situation de crise, les opérateurs à mobiliser et la mise en place d'une gouvernance immédiatement opérationnelle. Par exemple, dans un contexte d'émergence virale, SpF pourrait avoir un rôle central à condition de disposer de moyens adaptés. Poser et valider le principe que la meilleure façon d'être prêt lors d'une crise consiste à avoir un système des données de santé riche et complet en temps normal (cf. 2 ci-dessus), et dont le fonctionnement et certains paramètres (par exemple, fréquence d'actualisation, accès) sont modifiables en période de crise.

4.2 Mettre en place, en lien avec tous les acteurs concernés, une cartographie des besoins prévisibles en données de santé en situation de crise et des analyses qui devront être réalisées sur ces données. Cette évaluation devrait se baser sur les expériences des crises sanitaires antérieures et un benchmarking international. Anticiper un renforcement lors d'une crise des systèmes d'information utilisés (notamment via des serveurs plus puissants et des moyens humains accrus).

4.3 Améliorer la coordination entre SpF et les Agences Régionales de Santé pour un recueil de données plus fluide et efficace lors des situations de crise.

4.4 Renforcer les partenariats entre agences de santé et des équipes de recherche et centres de référence spécifiques hors temps de crise pour améliorer la capacité à produire rapidement des analyses essentielles pour l'aide à la décision publique (par exemple pouvoir rapidement analyser les données françaises avec des modélisations statistiques et mathématiques). Cela nécessite des mécanismes pour faciliter la mise à disposition des données ainsi que des financements appropriés.

4.5 Intégrer la réflexion sur les données de santé, les SI et leur interconnexion dans les différents plans de préparation et de lutte contre les émergences, y attacher de réels moyens financiers et prévoir leur mobilisation rapide et efficace.

4.6 Mettre en place dès maintenant un cadre réglementaire spécifique et des processus dérogatoires en situation de crise qui permettraient un accès facilité des chercheurs aux données de santé et la réutilisation des données déjà existantes et collectées dans le cadre du soin, en anticipant les autorisations règlementaires nécessaires.

4.7 Anticiper le recueil de données de surveillance et de recherche en situation de crise en mettant en place :

- **des contrats-cadre prédéfinis** ainsi que des **protocoles-type** d'étude de recherche pré-approuvés par les comités réglementaires, qui pourraient être enclenchés et financés rapidement en cas d'émergence. Ceci implique de disposer des moyens humains et infrastructures mobilisables rapidement.
- **des mécanismes rapides de financement et de déploiement de travaux de recherche spécifiques** et complémentaires de la surveillance en période de crise.

4.8 **Évaluer** les dispositifs mis en place par la réalisation de **tests de scénario de crise** sanitaire ou exceptionnelle en lien avec les experts des thématiques.

5 Impliquer davantage les citoyens dans le processus de prise de décision concernant les données de santé

5.1 **Renforcer l'implication des citoyens dans les processus de gouvernance des données de santé**

5.2 **Rendre plus efficaces les modalités de consentement et d'information pour la réutilisation de données de santé en situation de crise sanitaire**.

5.3 **Généraliser l'implication des personnes concernées dans les projets de recherche** et faire en sorte que la recherche soit davantage participative (participation aux comités scientifiques, à l'élaboration des protocoles, à l'analyse des résultats, à la valorisation, etc.).

5.4 **Communiquer de façon transparente :** Informer clairement le public sur les risques potentiels et les bénéfices du partage de données de santé, notamment en matière de recherche et pour la décision publique. Assurer la transparence des processus de collecte, d'utilisation et de partage des données de santé, ainsi que la reddition de comptes des organisations impliquées dans ces activités envers le public. Soutenir davantage la communication scientifique et sa diffusion à un large public.

5.5 **Faciliter les possibilités d'exercer son droit de retrait ou d'opposition via par exemple un portail centralisé.**

5.6 **Encourager les retours d'expérience** et les commentaires des citoyens sur les politiques et les pratiques de partage de données de santé, et **s'engager dans un dialogue continu** pour répondre à leurs préoccupations et améliorer les processus.

5.7 **Mettre en place une réflexion éthique à l'échelle de la société** sur les évolutions possibles des SI, leur interopérabilité et accessibilité, dans une vision à long terme intégrant le développement de l'Espace européen des données de santé.

6 Innover dans la génération de données de santé en anticipation des crises sanitaires

6.1 **Soutenir les approches innovantes en matière de surveillance** et de collecte de données utiles pour la surveillance et la recherche (designs d'étude et de surveillance innovants, réseaux sociaux, données de mobilité, eaux usées, nouvelles approches microbiologiques, utilisation de l'IA pour la surveillance...).

6.2 **Soutenir le financement public de la recherche en biologie-santé-environnement**, qui a fortement diminué au cours des deux dernières décennies, et soutenir les efforts en matière de science ouverte. **Soutenir fortement la recherche est une condition nécessaire pour disposer d'une expertise solide et d'outils sur un sujet donné lors d'une crise sanitaire future.**

I- Introduction – les données de santé

A. Données de santé : de quoi parle-t-on ?

L'expression de donnée de santé est actuellement utilisée avec différentes acceptions, ce qui ajoute de la complexité aux questions autour de leur production, analyse et partage.

1- Définition selon le contenu

Sur le plan juridique, la notion de « données de santé » est relativement récente : abordée pour la première fois seulement en 1995 par la CJCE dans l'arrêt Bodil Lindqvist (*CJCE, 6 nov. 2003, aff. C-101/01, Lindqvist*[4]), puis clarifiée par la CNIL et la jurisprudence sous le régime de l'ancienne loi informatique et libertés[5], l'article 4 alinéa 15 du RGPD les définit comme toutes « 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 ». Les données de santé sont ainsi définies selon leur contenu et la définition du RGPD peut être qualifiée de définition "large" des données de santé. L'expression "données de santé" est également largement reprise dans la loi française de 2019 relative à l'organisation et la transformation du système de santé et utilisée pour désigner les données incluses dans le Système national des données de santé (SNDS), qui inclut une base principale (elle-même issue du regroupement de plusieurs bases à caractère national) et un catalogue de bases (études, registres…) ayant souvent une couverture spatiale et temporelle plus restreinte.

Les données de santé recouvrent communément les données médicales relatives aux déterminants généraux de santé, les données individuelles sur l'état de santé d'une personne ou encore les données populationnelles de santé publique, de santé au travail ou encore de santé reproductive. Elles peuvent prendre la forme d'informations administratives, découler d'examens médicaux ou encore être le résultat d'un croisement de plusieurs bases de données. A ce titre, la CNIL distingue les données de santé en trois catégories : celles qui sont des données de santé par nature (maladies, antécédents médicaux, handicap, etc.), celles qui le deviennent du fait de leur croisement avec d'autres données, car elles permettent d'obtenir des renseignements sur l'état de santé d'une personne, et enfin, celles qui deviennent des données de santé en raison de leur destination[6]. Toutes les définitions s'accordent sur le fait qu'il s'agit de données personnelles (par exemple, des données agrégées sur l'incidence d'une pathologie à l'échelle d'une ville ou d'un pays ne seront pas comprises comme des données de santé, bien qu'il s'agisse de données sur la santé). Il faut enfin signaler d'autres données ne relevant pas de la santé humaine mais utiles en cas de crise (mobilité, exposition environnementale, santé animale...).

2- Définition selon la source

Les données de santé au sens large peuvent également être distinguées selon leur source (Figure 1).

[4] « l'indication du fait qu'une personne s'est blessée au pied et est en congé de maladie partiel constitue une donnée à caractère personnel relative à la santé ».
[5] Etaient alors considérées comme des données de santé « santé les données relatives aux addictions, à la dépendance, aux motivations à l'arrêt du tabagisme par exemple » selon Délib. CNIL n° 2013-282, 10 oct. 2013.
[6] https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante

Figure 1 : Les 4 grands types de données personnelles sur la santé, et le cadre règlementaire correspondant.

3- *Définition selon la finalité*

Enfin, on peut aussi distinguer ces données selon leur finalité, qui peut être la production de connaissances en recherche, la surveillance, la pharmacoépidémiologie, la mise en œuvre et l'évaluation des politiques de santé.... La recherche concerne typiquement les questions nouvelles, par exemple de nature étiologique ou mécanistique (le facteur X influence-t-il la survenue de la maladie Y ? Quel est le mécanisme physiopathologique de survenue d'une pathologie donnée ?). Cette vision recoupe partiellement la précédente, mais pas totalement, dans le sens où des données de recherche peuvent être utilisées en surveillance (par exemple si on utilise une cohorte à couverture nationale pour estimer la fréquence d'un facteur de risque connu), et symétriquement (par exemple lorsque des données de surveillance viennent enrichir des données de recherche).

Il faut noter que, si l'expression de donnée de santé est relativement récente, le recueil et l'analyse de données de santé sont beaucoup plus anciens et se sont organisés avec différents objectifs (soin, assurance, recherche, surveillance…). **Ces domaines se sont structurés selon des logiques et finalités différentes, et en silo, sans volonté d'ouverture et de partage de données**. L'apparition d'une thématique des données de santé fait apparaître une nouvelle logique, plus transversale, qui percute cette logique du passé et pose la question de l'optimisation de l'organisation des différents acteurs dans la collecte et l'analyse des données de santé.

B. Diversité des dispositifs de collecte de données de santé

Il existe une grande diversité des dispositifs de collecte de données, qui en retour donne lieu à la production de données de santé de source, de nature et de finalité différentes :

1) Le SNDS,

Le SNDS a été élargi par la loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé sur la base d'un postulat d'universalisation : tous les actes et prestations financés par la collectivité ont vocation à relever du SNDS. Le SNDS comprend notamment des données de registres, des cohortes de recherche, des entrepôts de données hospitalières, etc. : autant de données relatives aux différentes étapes des parcours de soins qui sont essentielles car complémentaires de la base principale du SNDS. Le SNDS a l'avantage de cibler

toute la population française (pas de biais de sélection ni de perdus de vue), de contenir des données individuelles médicalisées, structurées et codées de manière standardisée et qui sont rapidement exploitables, récentes et actualisées.

2) Les données de surveillance spécifique et de surveillance non spécifique

Ces données sont issues par exemple des systèmes de déclaration obligatoire (ex : Maladies à déclaration obligatoire, MDO) *et de surveillance non spécifique* (ex : services d'urgence, SOS médecins). Ici, la source SURSAUD (surveillance syndromique des urgences et des décès), créé en 2004, est une source importante qui a notamment été utilisée pour le Chikungunya (2005-2007), la rougeole, les attentats de 2015, la tempête IRMA de 2017 ou encore le Covid-19 en 2020. Elle comprend 4 sources complémentaires : les urgences hospitalières (OSCOUR, avec 94% des passages nationaux – 75% de passages avec un diagnostic médical codé) ; SOS médecins (95% des actes– 95% des actes avec un diagnostic codé) ***;*** la mortalité toutes causes issues des bureaux d'état-civil (84% de la mortalité nationale) ; les certificats électroniques incluant les causes médicales de décès (37% de la mortalité nationale). On peut également citer le réseau Sentinelles qui suit et détecte les épidémies d'infections respiratoires aiguës, de gastroentérites et d'autres événements en médecine ambulatoire depuis 40 ans ; les centres de références... **A ce jour, aucune de ces données, pourtant essentielles, ne sont intégrées dans le SNDS.**

3) Les entrepôts de données de santé (EDS) interdisciplinaires

Ceux-ci agrègent et structurent les données produites à l'occasion du soin dans les établissements de santé, et permettent ainsi de dépasser les limites du modèle classique informatique hospitalier qui s'organise par logiciels métiers spécialisés. Les EDS, qui connaissent actuellement un développement rapide en France, représentent un véritable atout en matière de données cliniques. Ils ont le rôle d'assurer la structuration de l'information médicale des patients qui fréquentent un établissement de santé au sein d'une base de données unique. Ils servent à différentes finalités, qui vont de l'amélioration de la prise en charge des patients, l'orientation des décisions des établissements de soins et le développement de la connaissance médicale[7]. Néanmoins, ils ne sont pas harmonisés, contiennent beaucoup d'informations non structurées (comptes-rendus par exemple) et l'accès à ces données est souvent difficile.

4) Les registres de santé

Définis dans l'arrêté du 6 novembre 1995 comme les « recueils continus et exhaustifs de données nominatives intéressant un ou plusieurs événements de santé dans une population géographique définie, à des fins de recherche de santé publique », ils comprennent ***les registres populationnels***, qui ont la particularité d'avoir une exigence d'exhaustivité du recensement de cas au sein d'une population donnée, ainsi que ***les registre de pratiques***, qui recensent la réalisation de certains types d'actes de santé. Selon le HCSP dans son rapport « registres et données de santé : utilité et perspectives en santé publique » (14 septembre 2021), *les registres sont mal définis et hétérogènes dans leur ressource et leurs missions*. Bien que menacés actuellement de disparaitre, le HCSP estime qu'ils ont un *rôle insubstituable en raison de la qualité de la collecte multi source de données reposant sur un travail humain*. De plus, si les registres ont une place centrale en épidémiologie, le HCSP recommande d'*étendre leurs missions/les renforcer dans d'autres champs de la santé publique, et notamment en termes d'aide à la décision publique*.

[7] https://www.health-data-hub.fr/actualites/entrepots-de-donnees-de-sante-un-programme-de-formation-pedagogique-pour-comprendre-de

5) Autres données de surveillance

Les enquêtes comportementales (ex : baromètre de SPF) ou épidémiologiques (enquêtes flash hebdomadaires), les données issues des Centres Nationaux de Référence (CNR), des réseaux (Obépine) et consortium de surveillance et/ou recherche (tel Emergen), les données de laboratoires d'analyses médicales (tel CERBA, BIOGROUP, BIOMNIS), les données formatées et non structurées (comptes-rendus d'examens…) et d'imagerie médicale, les données de pharmacovigilance ou d'accès précoce (ANSM), les données de consommation de produits médicaux, les données de mobilités et d'objets connectés…

6) Les données de recherche en santé (cohortes, études cas-témoins, autres études observationnelles, essais randomisés contrôlés…)

Noter que les registres et les données de surveillance peuvent aussi être pertinentes pour la santé, et elles sont parfois générées dans le contexte de structures de recherche.

7) Autres données pertinentes pour la santé

Il s'agit notamment des données d'exposition environnementale issues d'agences ou d'organismes de surveillance, les données sociales - notamment celles produites par la statistique publique, les données de santé animale... chacune de ces sources de données étant organisée avec son propre dispositif de collecte.

C. L'enjeu essentiel du chaînage des données – Le NIR comme clé d'appariement

Les données de santé sont une source extrêmement riche d'informations pour étudier et comprendre les déterminants de la santé des populations, évaluer l'efficacité des traitements, des vaccins et des mesures de santé publique, informer les autorités et guider le développement de mesures de santé publique efficace. Toutefois, la puissance de ces analyses dépend fortement de la capacité à apparier les différentes bases de données disponibles.

Comme le montre l'analyse des systèmes d'informations mis en place pendant la pandémie de COVID-19 (paragraphe VI. A), le besoin de chaînage des bases n'a pas été pris en compte dans la phase de conception des bases ce qui a entraîné des retards dans la production d'informations essentielles. Ce manque d'anticipation est dommageable. Pour pouvoir chaîner sans erreur les données d'un même individu issues de différentes bases, chaque base doit contenir un même identifiant individuel unique.

Il existe différents identifiants :

- le **NIR ou Numéro d'Inscription au Répertoire** national d'identification des personnes physiques (RNIPP), plus communément connu comme "**numéro de sécurité sociale**". Pour une personne, le NIR de l'usager (ou du bénéficiaire) est celui qui identifie la personne. Il peut être différent du NIR de l'ouvrant droit ou de l'assuré, par exemple si l'usager est un enfant, et l'ouvrant droit un parent. Le NIR est attribué par l'INSEE pour les personnes nées en France (métropole et DROM) ou par la CNAV par délégation de l'INSEE pour les personnes nées à l'étranger et dans les collectivités outre-mer à travers un système miroir du RNIPP, le SNGI (Système National de Gestion des Identités)

- le NIR de l'usager est utilisé comme le matricule identifiant national de santé (INS), mais l'usage de la terminologie "INS" est réservé pour le référencement des données de santé à des fins de prise en charge sanitaire ou de suivi médico-social. On parle de matricule INS pour le NIR, et d'identité INS pour le matricule INS + les traits d'identité de référence.

- le **hNIR** (**NIR haché**) est l'identifiant obtenu après traitement cryptographique irréversible et secret du NIR qui sert de référencement **dans le SNDS**. Pour pouvoir procéder aux appariements avec le SNDS sur la base du NIR, il est nécessaire que les mêmes traitements (hachage, cryptage) soit appliqués au NIR de la base que l'on cherche à apparier. C'est aujourd'hui **la CNAM qui dispose de ces algorithmes secrets** et qui effectue cette opération de *pseudonymisation*.

- **Le Code statistique non signifiant (CSNS)**[8] a été défini par la loi pour une République numérique de 2016 afin de permettre la mise en œuvre d'appariements de fichiers à des fins statistiques en limitant l'usage du NIR, et garantir ainsi un niveau élevé de protection des données à caractère personnel. Il s'agit d'une **clé d'appariement calculée à partir d'un chiffrement irréversible du NIR**. Ce service est **offert par l'INSEE** depuis octobre 2021 (à partir du NIR) et octobre 2022 (à partir des traits d'identité permettant de générer le NIR), **réservé aux organismes du service statistique public** (Dares, Drees, SDES et SIES en plus de l'Insee) et s'applique aux fichiers administratifs ou issus d'enquêtes. **Le CSNS est donc différent du hNIR**.

- **les traits d'identité** (nom, prénom, date de naissance et lieu de naissance) lorsqu'ils sont collectés, peuvent permettre de retrouver/reconstruire le NIR. C'est **la CNAV (via interrogation du SNGI)** ou l'**INSEE (via RNIPP)** qui dans ce cas jouent le rôle de tiers de "reconstruction".

Concernant les données de :

- **santé,** le décret "Cadre NIR"[9] précise les catégories d'acteurs et les finalités de traitements autorisant le recueil du NIR (https://www.cnil.fr/fr/tout-savoir-sur-le-decret-cadre-nir-dans-le-champ-de-la-sante). L'usage est en particulier autorisé pour la remontée d'information vers les organismes d'assurance maladie (ATIH), la mise en œuvre du dossier pharmaceutique (Conseil National de l'ordre des pharmaciens), la mise en œuvre du dossier médical partagé (CNAMTS). On peut noter également que l'usage du NIR est autorisé pour le suivi des travailleurs exposés au rayonnement ionisants (IRSN) et pour la gestion et le suivi des alertes sanitaires (SpF).
- **recherches en santé,** les traitements de données de santé mis en œuvre aux fins de réalisation d'études, d'évaluations et de recherches et relevant des dispositions des articles 72 et suivants de la loi « Informatique et Libertés » **ne rentrent pas dans le champ d'application des dispositions du décret « cadre NIR »**. Dès lors que la collecte du NIR est nécessaire pour ces traitements**, l'organisme responsable de traitement doit justifier de la collecte de cette donnée dans la demande d'autorisation qu'il adresse à la CNIL préalablement à la mise en œuvre du traitement.** Le dossier de demande d'autorisation doit nécessairement comporter un schéma de circulation des données, une analyse d'impact relative à la protection des données (AIPD) détaillant notamment de manière précise les mesures de sécurité dont la mise en œuvre est envisagée.

Par ailleurs, **la loi "pour une République numérique"** de 2016 a aussi prévu un **dispositif spécifique pour le monde de la recherche**. Une même opération de chiffrement du numéro de sécurité sociale pour aboutir à un

[8] Yves-Laurent Benichou, Lionel Espinasse, Séverine Gilles. Le code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers. Courrier des Statistiques n 9. Juin 2023. INSEE, pages 64-85

[9] 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

code non signifiant utilisable comme clé d'appariement devient une nouvelle possibilité offerte aux chercheurs. La différence avec le CSNS est toutefois que ce code de recherche est attaché à un projet de recherche en particulier et ne peut pas être utilisé pour un autre projet. Ainsi un individu avec un code non signifiant dans un projet de recherche n'aura pas le même code dans un autre projet de recherche.

Il apparait donc, en l'état de la règlementation actuelle, que :

- le recueil du NIR ou de traits d'identité permettant de générer le NIR (via sollicitation de la CNAV ou INSEE) peuvent être organisés pour les données de la recherche mais nécessitent la mise en place de circuits déconnectés des données recueillies, impliquant des organismes tiers et une demande d'autorisation spécifique à la CNIL,

- les identifiants *pseudonymisés* comme le hNIR (pour l'appariement au SNDS) et le CSNS (statistique publique) ne sont pas "identiques" car gérés par des organismes différents (CNAM dans le premier cas, INSEE dans le second)

La plateforme nationale des donnés de santé (le HDH) prévoit le déploiement en 2024 d'un concentrateur pour industrialiser les appariements au SNDS en utilisant le NIR ou les traits d'identité. La fonction de ce concentrateur est notamment d'industrialiser l'interrogation du SNGI (via la CNAV) par génération du NIR à partir des traits d'identité.

Le COVARS considère qu'il convient :

- **d'harmoniser ou interconnecter NIR et CSNS** pour faciliter les appariements entre données du SNDS et données de la statistique publique ;
- **de simplifier considérablement le recueil du NIR** en recherche, et prévoir un cadre règlementaire (qui peut reposer sur des accords / conventions cadres avec des organismes identifiés) qui permette une autorisation expresse en cas de crise sanitaire. On pourrait envisager de mettre en place d'une méthodologie de référence spécifique ou d'intégrer la question du NIR dans les méthodologies de référence existantes ;
- **de suivre le déploiement et l'activité du Concentrateur** prévu par le HDH et ses partenaires ;
- **de systématiser le recueil d'une clé d'appariement** (qui peut être le NIR, le CSNS ou autre) pour toute recherche en santé en mettant en place le circuit de traitement de données qui préserve la sécurité et la confidentialité des données. Ceci nécessite d'identifier une infrastructure dotée de moyens dimensionnés qui assure ce rôle de tiers pour les responsables de traitement des projets de recherche.

II. Constats et pistes d'amélioration

A. Retour d'expérience sur la mise en place de bases de données sur le soin et de recherche au cours de la crise COVID-19

1- Bases de données sur le soin et la surveillance

Au début de la pandémie de COVID-19, il n'existait pas de systèmes d'information permettant de recueillir les données nécessaires au suivi de l'épidémie. Divers acteurs se sont donc mobilisés pour créer dans l'urgence des systèmes d'information ad hoc permettant de répondre aux besoins. Les bases principales disposaient chacune d'un opérateur différent : La Caisse Nationale d'Assurance maladie pour les données de vaccination Vaccin-Covid[10], le ministère de la Santé et l'AP-HP pour les données de dépistage SI-DEP, et la DGS pour les données d'hospitalisation SI-VIC. D'autres bases ont progressivement été ajoutées par d'autres acteurs comme EMERGEN pour les séquences virales.

Globalement, le système a fonctionné et a rempli son rôle, avec des évolutions notables au cours des premiers mois de l'épidémie. Toutefois, cette mise en place, montée dans l'urgence et peu coordonnée, a impliqué des solutions sous-optimales, beaucoup de créativité et, a finalement conduit à des difficultés importantes pour la collecte, la gestion, le chaînage et l'analyse des données, avec des effets délétères sur la capacité à produire rapidement les connaissances nécessaires pour guider la décision en santé publique.

Afin d'éviter que ces difficultés ne se reproduisent lors de crises sanitaires futures, le COVARS considère que les systèmes d'information qui seront utilisés lors de ces crises devraient dans la mesure du possible anticipés, être construits voire déployés dès aujourd'hui.

2- Génération de données de recherche sur la Covid-19

Le monde de la recherche, dans toutes ses composantes, a fait preuve d'une mobilisation importante dès le début de la crise COVID-19, pour développer des travaux dont la finalité était d'apporter des connaissances utiles pour la gestion et le contrôle de cette pandémie. Les procédures d'autorisations règlementaires accélérées et le soutien financier immédiat accordé aux projets -par exemple via les appels d'offre "flash" de l'Agence Nationale de la Recherche (ANR)- ont favorisé cet élan, en tout cas pour des projets de petite taille. Néanmoins, et malgré les efforts du consortium multidisciplinaire français REACTing de l'Inserm (Research and Action Targeting emerging infectious diseases - créé en 2013 pour coordonner la recherche sur les émergences infectieuses), la coordination et la priorisation de la recherche, en particulier de la recherche clinique, ont initialement été insuffisantes[11], entraînant une prolifération d'études, certains se retrouvant en concurrence les unes avec les autres ou ne disposant pas des ressources (budget, effectif) nécessaires pour répondre aux

[10] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000042739429

[11] Il faudra attendre fin 2020 la mise en place du "Capnet" (comité ad hoc de Pilotage National des essais thérapeutiques et autres recherches pour les essais thérapeutiques et autres recherches sur le COVID-19) sous tutelle du MESR et du Ministère de la santé et de la prévention, pour qu'une coordination nationale de la recherche clinique soit installée (P Rossignol, Mission essais cliniques en contexte épidémique, juin 2020, accessible à : https://sante.gouv.fr/IMG/pdf/rapport_mission_essais_cliniques_p_rossignol_07062020.pdf)

questions posées. Les moyens disponibles pour lancer de nouveaux grands projets de recherche ou surveillance se sont aussi révélés limités ou difficiles à mobiliser.

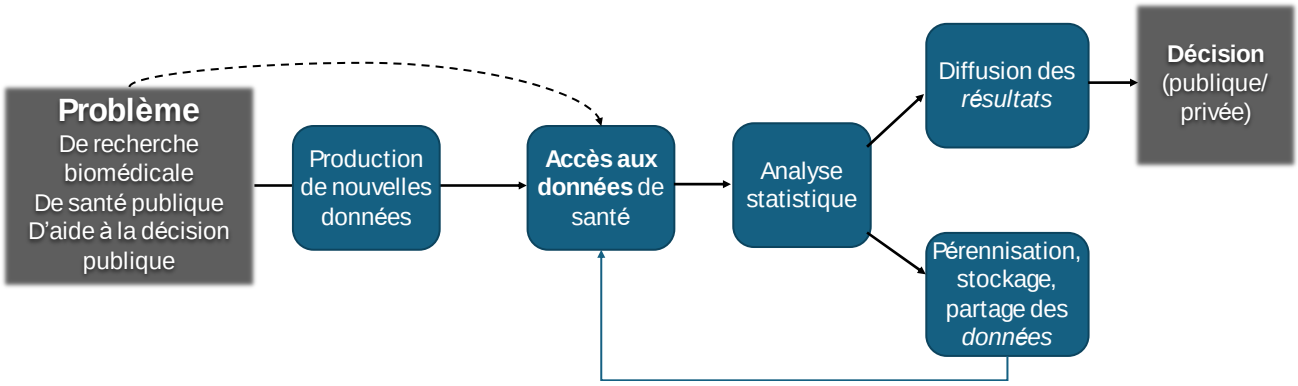
Deux autres facteurs ont pu être des éléments bloquants importants pour la production de données issues de la recherche pendant la crise COVID-19 : le manque d'infrastructures (ressources biologiques et plateaux techniques, personnels de recherche, ressources "data") mobilisables et en capacité de soutenir des recherches **en population ou en dehors de l'hôpital** alors que ces infrastructures existent au niveau des CHU et pouvaient être mobilisées ; des délais administratifs entre partenaires (accord de consortium, conventions) ou financiers (délais de mise à disposition des crédits, marchés publics) retardant l'exécution de ces recherches.

*Afin de répondre à une future crise sanitaire, **le COVARS insiste sur l'importance de garder à l'esprit que de nouvelles questions appellent généralement un fort besoin de générer de nouvelles données**. La question du partage des données existantes, qui ont aussi bien sûr leur importance, ne doit pas être traitée indépendamment de celle de la génération de nouvelles études et données, avec qui elle constitue un continuum rappelé en*

Figure 2.

Le COVARS considère qu'il est donc essentiel de mettre en place les mécanismes permettant l'identification rapide de grandes questions, puis de pilotes opérationnels dans le monde de la surveillance, recherche et soin pour chacune d'entre elles, assorties de moyens conséquents alloués rapidement à ces pilotes.

Figure 2 : L'accès aux données de santé n'est qu'une des étapes du cheminement allant d'un besoin ou question de santé, recherche, innovation à une décision ; à ce titre, elle ne doit pas être trop séparée de l'ensemble de la chaîne, qui débute par la formalisation du problème et la réalisation d'études et expérimentations éventuelles.



D'autre part, **les frontières entre questions de surveillance, statistique publique et recherche tendant à s'estomper en contexte de crise,** ceci implique, en termes de pilotage, que, si ces grands champs peuvent être pilotés dans une logique de relative indépendance, avec une instance de coordination, il peut être pertinent de revoir et potentiellement resserrer la gouvernance en période de crise ; ceci n'implique pas forcément un

pilotage fort et unique pour les « données de santé » en période de crise, d'autres modalités de structuration autour de grandes questions (incidence et facteurs de risque en population générale, tests, vaccination, modélisation et grandes interventions, recherche clinique sur les traitements, suivi des patients, par exemple) pouvant être adoptées, chacune requérant une composante sur les données de santé qui peuvent faire avancer la question.

Sont détaillées ci-dessous plusieurs difficultés apparues durant la pandémie de COVID et d'autres épidémies, ainsi que les solutions proposées à mettre en place.

B. Des systèmes d'information *ad hoc* communiquant mal entre eux et alourdissant la charge de travail

1- Pendant la pandémie de Covid-19

Les systèmes d'information tels que SI-VIC et SI-DEP n'avaient pas initialement été pensés pour absorber la masse de données qui a été recueillie durant la pandémie de COVID-19. Par exemple, le système SI-VIC de surveillance des hospitalisations a été conçu pour le suivi des victimes d'attentats et de situations sanitaires exceptionnelles. Quant à SI-DEP qui monitorait l'administration des tests, il s'agit d'une solution créée dans l'urgence par l'Assistance Publique-Hôpitaux de Paris et qui a ensuite été déployée à l'ensemble du territoire national. Ces systèmes d'information ad hoc ont nécessité un effort important des acteurs comme les hôpitaux et laboratoires. En effet, l'absence d'outils de communication entre les systèmes d'information des opérateurs de ces bases et des acteurs peut forcer ces derniers à ressaisir leurs données dans la base principale. Ainsi, alors que les hôpitaux ont connu une pression très forte durant la pandémie de COVID-19, ils ont dû dégager des ressources humaines pour renseigner quotidiennement la base SI-VIC. Cette situation peut donc représenter un coût RH substantiel pour les acteurs ; et la lassitude ou le manque de moyens peut conduire à leur désengagement, remettant en question la qualité des données. Cela a pu être observé à plusieurs moments pendant la pandémie de COVID-19, par exemple durant les vacances ou les périodes de décrue de l'épidémie. Les acteurs ont globalement fourni l'effort demandé pour faire face à cette crise sans précédent. Cependant, dans d'autres situations jugées moins critiques, ils pourraient ne pas être en capacité de déployer les ressources nécessaires, avec des effets délétères sur la qualité et la pérennité de la surveillance.

2- Lors de l'épidémie de Mpox

Plusieurs agences peuvent contribuer au recueil et à l'analyse des données épidémiologiques durant les épidémies, ce qui peut conduire à des difficultés pour le suivi et la gestion de ces épidémies. Par exemple, lors de l'épidémie de *Monkeypox (Mpox)*, les Agences Régionales de Santé (ARS) conduisaient les investigations épidémiologiques sur le terrain pour identifier les cas et mettre en place les mesures de contrôle locales, alors que SpF suivait la situation au niveau national. Des systèmes d'information communiquant mal entre les niveaux régionaux et nationaux pouvaient forcer les personnels des ARS, parfois débordés, à renseigner plusieurs fois les mêmes informations dans des bases différentes. Ainsi, l'absence de systèmes d'information communiquant peut dégrader la qualité de l'information disponible au niveau national et créer des tensions entre le national et le régional, concernant l'allocation des ressources humaines aux différentes missions des ARS.

Afin d'assurer la qualité et la pérennité des systèmes de surveillance, le COVARS considère essentiel de :

- **maximiser l'adhésion des acteurs** (hôpitaux, laboratoires) **qui vont participer à la surveillance. Pour cela, il faut autant que possible construire des systèmes d'information qui minimisent les coûts pour ces acteurs** (associés par exemple à la ressaisie des données), **tout en offrant de nouvelles fonctionnalités utiles à leurs activités.**
- **construire ces systèmes d'information en période inter-crise.** Le temps de la crise n'est pas le bon moment pour cela.

Exemple du système de surveillance des données de laboratoire (MIBA) du au Danemark

Ce système mis en place dès 2010, illustre qu'il est possible et bénéfique de construire un système d'information pérenne utilisable en tant de crise et hors crise. Le MIBA contient tous les résultats de tests microbiologiques des hôpitaux publics et privés danois ainsi que des données cliniques de patients. Une des forces du système est que les données sont automatiquement extraites des bases des laboratoires si bien qu'il n'y a pas de double saisie ni de délai de déclaration. Les résultats des tests sont transmis automatiquement et en temps réel à MIBA, qui est mise à jour toutes les 15 minutes. MIBA est intégrée directement dans les dossiers électroniques des patients et fait partie intégrante des soins de santé des patients. Lorsqu'un patient est testé, le résultat des analyses de laboratoire est à la fois injecté dans le système de surveillance, renvoyé vers le médecin qui a prescrit le test et intégré dans le dossier médical électronique du patient. Ce dernier y a accès via un portail public. Ainsi, MIBA permet au système de surveillance national d'être basé sur des résultats électroniques de laboratoire en temps réel. Il constitue également un accès national aux résultats de tests pour les cliniciens et les patients et une source de données pour la recherche. MIBA dispose également d'un transfert automatique de données vers d'autres bases de données telles que les bases de données cliniques. Les divers interfaces utilisateurs (pour les laboratoires, pour les cliniciens et les patients) offrent une plus-value importante qui motive les acteurs. Grâce à ce système mis en place en 2010, le Danemark disposait d'un système d'information extrêmement performant dès le début de la pandémie. Les Danois n'ont eu qu'à le redimensionner pour assurer un suivi détaillé des tests tout au long de la crise.

Le COVARS constate que :

- **La qualité des données recueillies durant l'épidémie de Mpox est décevante et interroge sur les leçons apprises de la crise COVID-19 et sur les rôles respectifs des opérateurs nationaux (SpF) et locaux (ARS) dans la production de données en situation d'émergence.**
- **Il est nécessaire et urgent de clarifier les organisations et la gouvernance pour le recueil de données durant les situations d'émergence et de développer des systèmes d'information plus performants** permettant de recueillir des données de qualité, homogènes entre régions et qui bénéficient autant à l'exercice des missions régionales qu'à la production de données nationales. Ceci peut passer par exemple par la mise à disposition de nouveaux outils : dashboards présentant de façon synthétique les données régionales, prévisions « nowcasting » de l'état actuel d'une épidémie en cours (évaluant le nombre de cas non détectés du fait des délais de déclarations ou de la sous-déclaration), évaluation de l'impact des mesures mises en place localement, prévisions du devenir local de l'épidémie.
- **Le développement de ces outils nécessite la mise en place de partenariats étroits entre Santé Publique France, les ARS et des équipes de recherche spécialisées dans ce type d'analyses.**

C. L'absence de clé d'appariement retarde l'interconnexion de différentes sources de données et limite la capacité de production de résultats importants d'évaluation des politiques de santé publique (en particulier vaccination)

1- Données de vaccination anti-Covid-19 : comparaison internationale

En France, le besoin de chaînage des bases n'a pas été pris en compte dans la phase de conception des bases, étant donné la situation de grande urgence dans laquelle les bases principales COVID-19 ont été créées, par différents acteurs et avec peu de coordination. En conséquence, les opérateurs ont eu le plus grand mal à chaîner les bases VAC-SI, SI-DEP, SI-VIC car il n'existait pas de clé d'appariement commune dans ces différentes sources de données (telle qu'un NIR de qualité). SpF a par exemple dû faire appel à un sous-traitant privé. Tout cela a conduit à des retards importants dans l'estimation de l'efficacité vaccinale, un élément pourtant indispensable pour guider les décideurs et informer la population. On notera également que SpF et la DREES ont fait leurs chaînages de façon indépendante, avec des algorithmes différents, conduisant à des bases chaînées différentes, et sans doute une duplication inutile des efforts.
Plus tard, l'intégration de SI-DEP dans le SNDS a été rendue complexe par un défaut d'anticipation dans le recueil du NIR des personnes testées. Le taux d'appariement spontané n'a pas dépassé les 50% pour les données telles que réceptionnées dans le SNDS (ce problème concernant à plus forte intensité les enfants). Ainsi, afin d'intégrer SI-DEP définitivement dans le SNDS, un travail a été réalisé par la DREES qui utilise des traits d'identité communs entre SI-DEP et SI-VAC, et l'appariement de SI-DEP au SNDS n'a été finalisé qu'en janvier 2024.

Ces difficultés d'appariement ont retardé la production d'estimations de l'efficacité vaccinale en France. **Au Royaume Uni,** le programme de vaccination a débuté le 8 décembre 2020 avec le vaccin ARNm Pfizer/BioNTech, à destination dans un premier temps des personnes âgées de plus de 80 ans et du personnel de santé. Dès le lancement du programme, Public Health England a mesurée l'efficacité de ces vaccins sur l'infection, les formes

symptomatiques, l'hospitalisation et le décès, avec des premiers résultats parus dès le 22 février 2022[12].**En Israël,** les premiers résultats de deux études d'efficacité en vie réelle ont été publiés dès janvier 2021 (Centre Médical Sheba et *Clalit Health Services*)[13]. **En France**, ce ne sera que 5 mois plus tard, en mai 2021, que les premiers résultats de l'étude Epi-Phare seront connus, quand bien même la campagne de vaccination a été lancée durant la même période qu'en Israël et au Royaume-Uni (27 décembre 2020)[14].

2- Données de la campagne de vaccination anti-Papillomavirus

Même hors situation exceptionnelle du COVID, des problèmes importants d'appariement de bases existent, créant des difficultés importantes pour l'évaluation des politiques de santé publique. Considérons par exemple la campagne nationale de vaccination HPV lancée en 2023 pour l'ensemble des collégiens de 5ème (11 à 14 ans). Ceux-ci pouvaient être vaccinés en situation ambulatoire (médecin, pharmacien, infirmier, sage-femme ...) ou au sein du collège par des équipes mobiles issues des centres de vaccination. Les données collectées dans cette deuxième situation qui ne donnaient pas lieu à un remboursement, étaient minimalistes et ne peuvent être intégrées dans le SNDS faute de recueil d'une clé d'appariement (NIR). De ce fait, les données relatives à la vaccination sont manquantes dans le SNDS pour environ 30% des collégiens vaccinés. Cela limite la capacité de réalisation d'études d'efficacité ou de sécurité liées à l'utilisation de ces vaccins. Dans un contexte d'hésitation vaccinale importante, l'absence de données de qualité permettant d'évaluer de façon fine l'impact des campagnes de vaccination est un enjeu de santé publique majeur et il est essentiel d'y remédier.

3- Freins à la recherche vaccinale contre les infections respiratoires

Un autre facteur intervient dans la difficulté d'évaluer l'efficacité et la sécurité des vaccins au fil de l'eau : il n'existe pas à ce jour de système d'information dédié qui retrace l'histoire de la vaccination d'un individu. Dans le projet **RESPIVAC**, qui vise à quantifier les efficacités vaccinales contre les différents virus respiratoires, cette absence de système d'information dédié à la vaccination rend difficile la détermination du statut et de l'historique vaccinal des patients, entraînant une perte de qualité des données.

Le COVARS considère qu'il est essentiel :

- **qu'un NIR de qualité soit disponible dans toute les bases afin de permettre un chaînage facile des bases existantes.**
- **de développer un système d'information dédié à la vaccination pour pouvoir suivre de façon fine l'efficacité et la sécurité des vaccins utilisés en France.**

Le COVARS soutient donc pleinement l'initiative de la DGS de créer un système d'information sur la vaccination.

[12] Etude Sarscov2 Immunity & REinfection EvaluatioN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3790399
[13] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00448-7/fulltext
https://www.nejm.org/doi/full/10.1056/NEJMoa2101765
[14] https://www.epi-phare.fr/rapports-detudes-et-publications/impact_vaccination_covid/

D. Difficultés d'accès des chercheurs aux bases existantes retardant des projets stratégiques

Devant l'émergence d'une pathologie nouvelle qui conduit à une crise sanitaire, il est essentiel de permettre une acquisition rapide de connaissances solides afin de faciliter les décisions de santé publique, ce qui impose de faciliter l'accès réglementé des chercheurs aux bases de données.

Dans un esprit de science ouverte, Santé publique France a déployé des efforts importants pour que les données de surveillance en santé agrégées (comptage des hospitalisations, dépistages et vaccinations, issus des systèmes SI-DEP, VAC-SI ; accessible via GEODES) soient mises en ligne en quasi-temps réel, avec un niveau d'agrégation suffisant pour garantir l'anonymat des patients. Cette initiative a permis à de nombreux scientifiques de produire des résultats scientifiques utiles notamment pour la modélisation.

En revanche, certaines analyses nécessitent l'accès sécurisé aux données individuelles pseudo-anonymisées, par exemple quand il s'agit d'étudier l'efficacité vaccinale ou des facteurs de risque d'infection. Dans ce deuxième cas de figure, force est de constater que la communauté scientifique a eu beaucoup de mal à accéder aux données, ce qui a pu retarder des projets pourtant stratégiques et importants pour la réponse. On peut citer les exemples suivants survenus durant la pandémie de COVID-19 :

- *Des difficultés rencontrées par le GIS EPI-PHARE* alors en charge d'étudier – entre-autres – l'utilisation des produits de santé, le risque de forme grave et l'efficacité/sureté des vaccins[15] : une absence d'accès à SI-VIC, un accès à SI-DEP via la CNAM qui nécessitait ensuite pour EPI-PHARE de réaliser lui-même le chaînage ; un accès à SI-DEP qui différait de celui d'autres organismes (dont SPF qui n'avait pas les mêmes données) ; des retards sur les données d'hospitalisation, ce qui a eu des conséquences négatives sur l'évaluation de l'efficacité vaccinale…[16]
- *L'estimation des taux d'hospitalisation et de la létalité Covid-19,* qui a nécessité l'analyse conjointe des données de cohorte SAPRIS-SERO pour les infections de la première vague et des données de surveillance SI-VIC (via SPF), a été complexifiée par la granularité insuffisante des données ouvertes en ligne.
- Les parcours d'accès aux bases peuvent être particulièrement longs. Cela a notamment retardé le travail du projet COV-POPART, qui visait à caractériser la réponse immunitaire après vaccination et sa persistance, ainsi que les échecs vaccinaux. Ce projet n'a pu débuter qu'en décembre 2020 alors que de nombreux centres avaient déjà commencé le même type de travaux de façon monocentrique. Pour ce même projet, l'accès au SNDS (chainage via le NIR), prévu dans le protocole de départ de 2021, n'avait pas encore été réalisé en janvier 2024.

Il est à noter que dans des pays voisins comme le Royaume Uni, les équipes de recherche ont eu moins de difficultés pour rapidement accéder aux données individualisées pseudo-anonymisées. Dans un contexte de compétition intense entre équipes de recherche, cela leur a donné un avantage important, qui se traduit notamment en termes de nombre et impact des publications scientifiques et d'informations disponibles dans le contexte français.

De façon générale, l'un des principaux freins à l'utilisation secondaire des données de santé réside aujourd'hui en les **délais d'accès aux données.** Au 1er janvier 2024, la CNAM est le seul organisme qui héberge les données du SNDS et ne dispose pas de moyens dimensionnés par rapport à la demande d'appariement de ces données. Ceci se traduit par une mise à disposition des données appariées par la CNAM qui peut prendre entre 10 et 12 mois, en sus des délais d'obtention des autorisations CESREES et CNIL, ce qui porte le délai global à 18 mois, un

[15] Par exemple, étude de cohorte sur 66 millions de personnes sur une cinquantaine de pathologies chroniques : https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(21)00135-6/fulltext

[16] Eléments rapportés par l'ANSM GIS EPI-PHARE lors de son audition par le COVARS du 3 juillet 2023

délai fortement préjudiciable à la recherche française. Par ailleurs il faut rappeler que les données accessibles ne sont pas mises à jour en temps réel (délai de disponibilité de plus d'un an pour les données consolidées du PMSI, de l'ordre de 2 ans pour le CEPIDC).

Le COVARS considère qu'en prévision de futures émergences, il est essentiel de :

- **Développer dès aujourd'hui, des mécanismes simplifiés et rapides pour que les scientifiques impliqués dans la réponse à ces émergences puissent rapidement accéder et analyser les données existantes.**
- **Réduire les délais d'appariement aux données du SNDS et ses délais de mise à jour. Cela passe par une mise à l'échelle moyens humains et technico-règlementaires des organismes en charge de ces accès et des appariements, des processus d'autorisation anticipés et simplifiés, des délais administratifs réduits.**

Cela passe aussi par une attention particulière à porter aux contraintes administratives et règlementaires, et à la coordination et la simplification des acteurs et de leurs relations, détaillées dans les deux points ci-dessous.

E. Contraintes administratives et réglementaires lourdes

Le RGPD prévoit l'information individuelle des patients ou des citoyens en cas de réutilisation secondaire de données. Du retour des associations de patients, il apparaît que les intéressés sont rarement informés de la réutilisation secondaire, ne savent pas où trouver l'information ni comment faire valoir leurs droits de retrait ou d'opposition. Cette clarification des obligations relatives à l'information des patients ou des citoyens est importante, car la multiplicité des démarches en fonction des cas rend difficile la compréhension des exigences réglementaires pour les utilisateurs.

Pour cela, il convient de :

- **Renforcer les moyens des organismes concernés et de la CNIL.**
- **D'organiser et simplifier l'obligation d'information individuelle en cas de réutilisation secondaire de données, par exemple en mettant en place un portail de transparence unique et facilement accessible.**

Du point de vue des agences et organismes impliqués, les principaux obstacles à l'accès aux données et à la production de connaissance sont administratifs, tels que les conventions entre organismes, les accords de transfert de données et les délais de recrutement de personnel spécialisé en analyse de données, qui représentent un frein majeur à la réalisation des recherches et à la compétitivité.

Le COVARS suggère donc de développer des procédures standardisées et simplifiées pour la négociation, la rédaction et la signature des conventions entre les différentes administrations. Cela pourrait passer par la création de modèles de conventions pré-approuvés pour des cas courants de réutilisation des données de santé. Il est important également d'établir des délais de traitement prédéfinis pour chaque étape du processus de mise en place de conventions juridiques entre les administrations, en veillant à ce qu'ils soient réalistes et respectés.

F. Multiplicité des acteurs, manque de vision stratégique et de coordination

De nombreux acteurs contribuent à la production, la gestion, l'analyse et l'utilisation des données de santé. La Figure 1 illustre les principaux producteurs de données de santé et leurs tutelles ministérielles.

Cette multiplicité des acteurs n'est pas forcément une mauvaise chose en soi, mais elle nécessite une vigilance particulière sur plusieurs aspects critiques comme la construction d'une **vision stratégique globale partagée** par tous les acteurs, s'appuyant sur :

- L'expertise des utilisateurs et portée politiquement,
- La bonne coordination entre acteurs,
- La qualité du chaînage entre les différentes bases des données.

Les questions liées aux données de santé peuvent rapidement devenir très techniques et il est donc important que de très bons techniciens de la donnée soient impliqués à toutes les étapes de la construction des systèmes d'information de demain. La contribution de ces techniciens est essentielle pour optimiser le design et l'implémentation des systèmes d'information.

En revanche, la réponse à des questions plus stratégiques (par exemple : Quels systèmes de surveillance innovants doit-on développer en priorité pour améliorer la surveillance et soutenir l'aide à la décision ? Quels investissements doivent être faits en priorité pour répondre aux questions de recherche de demain ?) doit être construite en premier lieu par les producteurs / utilisateurs, comme les agences de santé (par exemple SpF), les professionnels de Santé Publique, les Directions des Ministères, les chercheurs, épidémiologistes, cliniciens, les représentants des citoyens et associations de patients etc. Les agences de santé qui devraient être centrales dans la construction de cette réflexion, apparaissent souvent à la marge. Les chercheurs sont également très peu représentés.

Sans implication adéquate des producteurs/utilisateurs dans la construction de cette vision stratégique, le COVARS considère qu'i**l y a un risque important que les mauvais investissements soient faits, conduisant à se tromper de priorité.** Par exemple, la Plateforme des données de santé a passé beaucoup de temps à tenter de développer un « catalogue », qui inclura à terme un certain nombre de bases et cohortes dont beaucoup sont de petite taille, avec une couverture souvent bien inférieure à 1% de la population vivant en France et d'un intérêt stratégique très limité ; ces efforts, dans un contexte de moyens limités, ont forcément limité ceux faits pour développer la base principale du SNDS, qui inclut des bases avec une couverture bien plus vaste et est d'un intérêt stratégique incomparable.

Le COVARS considère sur la base de ces retours d'expérience qu'il est important de :

- **Construire une vision stratégique globale partagée par tous les acteurs de la donnée de santé,** s'appuyant sur l'expertise des utilisateurs et producteurs et portée politiquement.
- **Clarifier les rôles et prérogatives en matière de gestion stratégique et d'exploitation des données de santé de manière à éviter tout sujet de périmètre en situation de crise**.
- **Renforcer le dialogue entre les directions les plus impliquées dans les données de santé au sens large** (DGS, Direction du ministère de la Santé en charge de la prévention et des crises, et son bras armé, Santé publique France ; DREES ; DGRI et l'Inserm en tant que pilote de la recherche en santé).

En effet, un examen du rôle des différents acteurs ne permet pas d'identifier facilement la complémentarité des rôles d'institutions dans l'accompagnement et le pilotage des données du soin (HDH, DNS, CNAM, DREES). A l'heure actuelle, il existe plusieurs instances réunissant tout ou partie de ces directions, notamment le CASA (Comité des agences en santé, piloté par la DGS) et le Comité stratégique des données de santé. Il ne nous appartient pas de préciser comment renforcer leur action stratégique en faveur de la génération et l'utilisation des données de santé.

G. Structuration opérationnelle des données de la recherche en santé

Le champ de la recherche en santé a eu tendance à se structurer dans une logique de bas en haut, avec des fonctions support juridiques et SI relativement limitées, et par grandes communautés de recherche thématiques (celle des maladies infectieuses, cardiovasculaires, du cancer…). Le partage des données de cette recherche a toujours existé, de façon hétérogène entre communautés (typiquement plus répandu dans les domaines où les consortia sont vitaux, comme les maladies rares ou la génétique), mais principalement entre experts d'une même communauté et d'abord dans le cadre de partenariats de recherche permettant de partager simultanément expertise et données. Ceci est cohérent avec le fait que le niveau de documentation des données est souvent limité, faute de moyens, et que ces bases sont parfois très évolutives, rendant un partage des données sans contact avec les investigateurs risqué (en termes d'identification des variables pertinentes, connaissances des limites de l'étude…).

Aujourd'hui, beaucoup d'études sont ouvertes au partage des données via un contact individuel avec l'équipe et éventuellement, un accord scientifique (ou contrat dans le cas de partenariats avec le privé), et en réalisant les démarches règlementaires nécessaires du côté du responsable du traitement en accord avec le RGPD. Néanmoins ces démarches sont souvent initiées *a posteriori* alors qu'elles devraient être anticipées et organisées *a priori*, et elles nécessitent des compétences et moyens humains (en data management, en expertise juridique) insuffisamment disponibles au niveau des structures ou des organismes de recherches. En conséquence, les partages des données issues de la recherche en santé sont réduits, et les producteurs ou utilisateurs sont souvent découragés par les obstacles à affronter, ce qui se traduit par un déficit de valorisation scientifique et une perte de compétitivité en recherche. Les organismes de recherche ne sont par ailleurs pas en mesure de documenter finement le partage et l'usage secondaire de ces données.

Il est essentiel que le milieu de la recherche en santé soit davantage structuré de ce point de vue en :

- **Implémentant des standards de données en recherche** (tout en dotant les organismes de moyens capables de développer et soutenir ces standards de données) afin d'en faciliter l'interopérabilité de ces données avec les autres champs de la donnée en santé ;
- **Centralisant :**

- les métadonnées sur les bases de données, et les points et modalités d'accès à ces données (le projet de portail des métadonnées FReSH, piloté par l'Inserm et l'IReSP, devrait le permettre) ;

- les statistiques sur le partage des données ;

- **Harmonisant**, ou au moins rendant plus lisibles les logiques et si possible les démarches de partage, avec les partenaires publics et privés ; le développement de convention cadre ou de formulaire standardisé d'accord de transfert ou de partage de données qui ne serait plus à valider au "cas par cas"
- **Faisant davantage dialoguer** de façon structurée avec les autres grands champs de la donnée en santé (soin, surveillance, statistique publique).

Ces efforts doivent être faits en cohérence avec la politique de science ouverte promue par le MESR, qui a l'avantage de mettre en avant des concepts et principes essentiels liés à la qualité de la donnée, la réplicabilité de la science, et d'insister sur la finalité essentielle, à savoir le partage des connaissances plus que celle des données primaires, souvent bruitées et très difficiles à interpréter sans expertise à la fois en science des données et sur la thématique scientifique précise concernée. Ils doivent se faire sans rompre le continuum allant de la génération des questions scientifiques au recueil puis à l'analyse, au partage et éventuellement la réanalyse des données et enfin le partage des connaissances, qui reste structurant et efficace pour générer des connaissances rigoureuses.

Cela devra être accompagné d'un renforcement très important des compétences en droit de la donnée, SI et sciences des données et fonctions "support" administratives, afin de limiter les freins techniques et administratifs aux recueil, analyse et partage des données.

La communauté des données de recherche en santé devra être capable de parler d'une voix et dialoguer avec les acteurs impliqués dans les autres grands types de données sur la santé (surveillance, soin, statistique publique) afin de définir une stratégie générale et d'harmoniser certains efforts. Cela devrait passer par un renforcement du MESR sur ces questions, portées par très peu de personnes, et surtout de l'Inserm en tant que coordonnateur de la recherche en biologie-santé.

Tous ces efforts sont d'autant plus nécessaires dans le contexte de la future mise en place de l'espace européen des données de santé (EHDS).