Pour l'utilisation des données de santé en vie réelle en recherche clinique

Etat des lieux, analyse et recommandations

GT AIS/F-CRIN "L'évolution des méthodologies d'essais cliniques : nouveaux outils, nouveaux usages et conditions de recours"
Sous-groupe "Données"

GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE 2030

AGENCE DE L'INNOVATION EN SANTÉ

F.CRIN
FRENCH CLINICAL RESEARCH INFRASTRUCTURE NETWORK

Mai 2024

GT AIS/F-CRIN « L'évolution des méthodologies d'essais cliniques : nouveaux outils, nouveaux usages et conditions de recours » – Sous-groupe « Données »

Membres : Mathieu Robain ; Stéphanie Kervestin ; Olivier Demarcq ; Gil Lampe ; Marc Cuggia ; Manon de Fallois

en concertation avec Camille Schurtz et Vincent Diebolt, copilotes du GT AIS/F-CRIN

Des remerciements particuliers sont adressés aux relecteurs du rapport et en particulier :

- La Filière Intelligence Artificielle & Cancers (FIAC)
- Les Entreprises du médicament (Leem)
- L'Alliance pour la Recherche et l'Innovation des Industries de Santé (ARIIS)
- L’équipe du « Data and medical evidence center » du laboratoire « Roche France »

Préface

Les données de santé sont définies à l'article 4 du Règlement général sur la protection des données (RGPD) 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* ». Le RGPD n'opère pas de distinction selon le contexte de la collecte des données : soin, recherche, primaire, reprise. Toutefois, en fonction de ce contexte, des dispositions spécifiques, notamment issues du code de la santé publique ou des règlements européens (relatifs aux essais cliniques ou aux dispositifs médicaux), peuvent s'appliquer concernant les modalités d'information et de recueil du consentement ou d'une non opposition en vue de la participation à la recherche (par exemple si les données sont collectées dans le cadre d'une recherche impliquant la personne humaine).

Les données de santé constituent la brique élémentaire indispensable à tout processus d'évaluation des produits de santé pour décider de leur mise sur le marché, la prise en charge par la collectivité pour leur remboursement, mais aussi tout au long de leur cycle de vie. Il est donc essentiel de pouvoir garantir leur disponibilité, leur qualité et de définir leurs conditions d'exploitation.

Aujourd'hui les sources, la nature et l'origine des données de santé se sont démultipliées. Elles peuvent relever des activités de la prise en charge par les professionnels de santé, de protocoles de recherches (dont des essais cliniques), ou directement des patients, que ce soit des données médico-économiques, observationnelles numérisées ou non, épidémiologiques, ou de toute sources de données témoin de la prise en charge des patients. Cette révolution, qui a constitué le point de départ du développement de nouvelles méthodologies d'essais cliniques, est à la fois une chance, en permettant de disposer d'un capital d'informations sur l'évolution de l'état de santé et la prise en charge des patients, et un enjeu tant leur nombre et leur hétérogénéité rendent complexe leur utilisation.

Ce travail de réflexion collective et les recommandations qui y sont formulés s'inscrivent en prolongement et complémentarité, ciblant l'activité de recherche clinique, du Rapport rendu public le 19 janvier 2024 : « Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé » (Mission coordonnée par Jérôme Marchand-Arvier), du Livre blanc : « Données de santé artificielles : analyse et pistes de réflexion » coordonné par le Pr Stéphanie Allassonnière et le Dr Jean-Louis Fraysse (avril 2024) et des contributions récentes sur l'usage des données de santé.

Ce rapport constitue le 1er volet d'une réflexion globale sur l'évolution des méthodologies en recherche clinique, objet d'une initiative lancée en mars 2023 par l'Agence de l'innovation en santé (AIS) et F-CRIN, Infrastructure nationale de recherche clinique. Dans la mesure où la réflexion engagée par le collectif réuni s'inscrit sur la durée ce rapport n'est pas un document figé mais au contraire évolutif qui pourra être actualisé, modifié, complété suivant l'évolution du contexte et la prise en compte et l'impact des recommandations.

18 recommandations pour favoriser l'utilisation des données de santé en vie réelle en recherche clinique

Axe 1 – Référencer et documenter les bases de données existantes pouvant être exploitées dans l'évaluation de l'innovation en santé

- **RECO 1 : Identifier et référencer au niveau national les bases et entrepôts de données accessibles** avec une caractérisation orientée « utilisateurs/chercheurs ». Ces paramètres de caractérisation devront permettre d'apprécier le périmètre et les usages compatibles, la maturité des données, leur niveau de qualification, le degré d'interopérabilité et les principes d'accessibilité (gouvernance, valorisation, modalités) …
- **RECO 2** : Impliquer les producteurs et utilisateurs de bases et d'entrepôts de données dans la définition du cadre des « **critères minimums** » compatibles à une reconnaissance et prise en compte des projets menés, et dans la définition du « niveau de preuve scientifique et grades des recommandations selon la typologie des données » issus des principes de « l'Evidence-Based Medicine ». En vue de généraliser l'utilisation des données de vie réelle, analyser l'apport des projets menés sur ces bases et entrepôts de données de santé de qualité.
- **RECO 3 : Standardiser la gouvernance et les aspects règlementaires pour le partage de données** issues de plusieurs bases ou entrepôts de données de santé (EDS) afin de favoriser l'accès aux innovations.
- **RECO 4** : Par la **priorisation thématique**, les financements publics de la recherche clinique, type PHRC, doivent inciter au recours à de nouvelles méthodes mettant en œuvre des données de vie réelle, comme la construction de bras de contrôle externe ou l'émulation d'essais cliniques cibles. Une concertation associant l'Agence de l'innovation en santé, le ministère de la Santé, l'Agence nationale de la recherche devrait permettre de définir un cadre d'application pour assurer cette thématisation.

Axe 2 – Garantir la qualité, l'interopérabilité, la standardisation et la pertinence des données au regard de l'utilisation prévue

- **RECO 5 : Définir des critères d'appréciation de la qualité de la donnée et pouvoir vérifier en amont via une phase de tests préalables son utilisabilité au regard de l'utilisation visée.**
  Cette étape de vérification s'apprécie en fonction de trois critères :
  1. La maturité du cycle de vie de la donnée, dont le cycle d'actualisation

2. L'appréciation du jeu de données vis-à-vis de ses caractéristiques intrinsèques, et le niveau d'adéquation de sa qualité avec les différentes familles d'analyse pour laquelle il est destiné
3. L'exhaustivité et l'exactitude des métadonnées qui décrivent le jeu de données

S'appuyer sur un modèle de qualité des données de référence, reconnu par ses pairs, tels que ceux en développement par l'Agence Européenne du Médicament (DQF for EU medicines regulation), s'appuyant sur les dimensions de cohérence, fraîcheur, pertinence, fiabilité et d'exhaustivité pour décrire la qualité et l'utilité des données, permettant d'assurer une compréhension commune, harmonisée et clairement documentée de la qualité des données. Cette démarche d'évaluation de la qualité doit être menée de façon indépendante et standardisée.

Elle suppose également la poursuite d'expérimentations et de davantage de « preuve de concept » de mise à disposition de jeux de données anonymisées lors d'une phase de tests pour en vérifier l'utilisabilité. L'accès à ces « bacs à sable » se ferait de manière simplifiée au plan juridique.

- **RECO 6** : Assurer **la prise en compte des données de vie réelle pour l'accès à l'innovation :** Intégrer dans la conception en cours du label européen de qualité et d'utilité des données par la Commission Européenne, la mesure de la qualité de données de vie réelles utilisées dans le cadre d'un essai clinique pour des données auditables par les agences règlementaires pour les accès cliniques.

- **RECO 7 :** Mettre en œuvre une **standardisation syntaxique et sémantique compatible avec les standards internationaux** comme vecteur de qualité, permettant de spécialiser et classer les données pour un cadre d'utilisation. Les référentiels sémantiques ne sont pas assez partagés, acceptés et promus par les utilisateurs (société savantes, collège d'experts,). Les utilisateurs devraient être incités pour une implication plus importante sur ces sujets. Ceci permet à la fois d'offrir un niveau minimal de qualité et de proposer une logique de test pour mesurer les dimensions de qualité qui n'auraient pas déjà été assurées.

- **RECO 8 : Développer la recherche sur les méthodologies de mise en qualité des données** intégrant l'identification et la neutralisation des biais et l'évaluation du niveau qualité réel des données des données de vie réelle (EDS), Cohortes). Ces recherches pourraient bénéficier d'un financement sur appel d'offre/appel à projets spécifique pour permettre aux équipes expertes sur ces sujets de mener ces travaux. Il pourrait s'agir par exemple d'essais hybrides dont les bras témoins seraient composés de données du SNDS, d'EDS ou des données de santé artificielles.

Axe 3 – Assurer la reconnaissance et le partage de la valeur dans la collecte et l'utilisation des données

- **RECO 9** : En complément des travaux en cours au sein du Comité stratégique des données de santé, **établir au plus vite des principes d'échange et de tarification pour l'accès aux différentes typologies de bases de données avec les partenaires privés.** Ces modalités sont à construire avec les représentants de l'ensemble des parties impliquées (HDH, EDS, établissements de santé, offices de transfert de technologie, investisseurs, grands groupes) pour aboutir à des grands principes (échelle des coûts, cadre règlementaire simplifié) acceptés par tous. Une lisibilité et une transparence sur les conditions de partage seront un critère fort d'attractivité.

- **RECO 10 :** Afin de rémunérer le travail sur les données à leur juste valeur, établir pour les cas d'usage identifiés, des **principes de partage de la propriété intellectuelle entre producteur de données et utilisateur** prenant en compte l'apport intellectuel du producteur. Ces règles de partage sont nécessaires lorsque la contribution du producteur des données dans l'acquisition, la qualification, la mise en forme, la documentation et la mise à disposition est effective.

Axe 4 – Améliorer les modes et circuits d'information du patient

- **RECO 11** : Tout en s'assurant de l'implication des patients et de leurs représentants dans la conception des notices d'information jusqu'à la communication des résultats des études, **mise en place de notices d'information simplifiées harmonisées** mentionnant systématiquement la possibilité de réutiliser les données et l'adresse web du portail de transparence alimenté au fil de l'eau afin de favoriser les usages secondaires de ces données.

- **RECO 12** : L'information individuelle des patients lors de la constitution et de l'exploitation des données d'un EDS ou concernant la réutilisation de leurs données (quand un portail de transparence n'a pas été mis en place) pourrait s'appuyer, par exemple, sur l'Espace Numérique Santé sécurisé afin d'assurer aux usagers une information accessible et la capacité de s'opposer à leur utilisation. Cette solution facilitera le travail d'information des porteurs de projets sur données d'EDS. **L'enjeu est aujourd'hui de faciliter le partage des données tout en respectant le choix des individus.**

  Pour cela, le portail monespacesante.fr porté par le ministère de la Santé et la Caisse nationale de l'assurance maladie pourrait être utilisé comme support privilégié d'information et de gestion de l'opposition à la réutilisation des données. Ce système centralisé permettrait de renforcer l'autonomie des patients et de garantir une meilleure gestion et utilisation des données.

Axe 5 – Inciter à la collaboration et au partage de données

- **RECO 13 : Définir des modalités de recevabilité précises par les Autorités des données/sources de données** (ex : données répondant aux normes de qualité de niveau règlementaire garantissant ainsi une « auditabilité » de ces données) en amont d'une évaluation, en vue de garantir une meilleure reconnaissance de la valeur des données de vie réelle utilisées dans le cadre de la recherche clinique. Afin de renforcer l'attractivité au plan international de ces sources de données, définir des conditions d'accès accélérées quand celles-ci ont pour objectif d'être utilisées dans le cadre de la recherche clinique. Des retours positifs sur la prise en compte des données issues des bases de données émis par les autorités de santé françaises, européennes et internationales seraient considérés comme une dimension du label de qualité.

- **RECO 14** : En lien avec le Comité stratégique des données de santé, **mise en place d'une instance dédiée à la construction de la doctrine relative à la constitution des bases de données.** Sa composition et sa gouvernance garantiraient une bonne représentativité de l'ensemble des acteurs (producteurs de données et monde socio-économique). Elle serait chargée de :

  - Définir une stratégie opérationnelle de données de santé en France basée sur l'usage et émettre des recommandations sur la constitution future des bases/EDS en adéquation avec les enjeux actuels et futurs de la recherche publique et privée (ex : France innovation santé 2030).
  - Assurer plus de cohérence entre les différentes sources (notamment accès précoce, registre, cohorte, génomiques…) et les usages de données de santé.

- **RECO 15 : Proposer un accompagnement personnalisé associant les régulateurs aux porteurs de projets souhaitant mettre en œuvre des essais innovants.** Sont directement concernés la CNRIPH, CNIL, HDH, ANSM. Le groupe de travail (GT) copiloté par l'AIS et F-CRIN va contribuer, en tant qu'amorçage et pré figuration, à définir les besoins et les modalités de réponse à y apporter.

  Ce GT serait chargé de publier des recommandations pratiques au fil des cas d'usage, à l'instar de ce qui est prévu dans le cadre de la phase pilote sur les « essais décentralisés et dématérialisés ». Par ailleurs, ces réflexions pourraient utilement nourrir les travaux de la CNIL dans la cadre de la mise à jour de ses méthodologies de référence.

- **RECO 16 : Définir un modèle économique** (mobilisant des soutiens publics et privés) assurant le financement de la création de bases de données et la collecte de données exhaustives et de qualité tout en définissant de nouveaux indicateurs de mesure de cette qualité.

  Ce modèle prendra en compte la reconnaissance de l'ensemble des acteurs ayant participé au travail de collecte et de validation des données dans la publication/communication des résultats.

Axe 6 – Assurer la dissémination des modes d'utilisation des données de santé en vie réelle en recherche clinique

- **RECO 17 : Informer et former les membres des organismes amenés à valider le recours à des données de vie réelle dans le cadre des protocoles de recherche clinique**. Le cadre de la formation existe aujourd'hui avec le partenariat et la plateforme mis en place par la CNRIPH, la DGS et l'EHESP.
- **RECO 18 :** Au sein des formations diplômantes en lien avec la recherche clinique et la recherche sur les données (Master en recherche clinique ; Master en biométrie ; Master en Datascience ; …) intégrer **des modules spécifiques sur les méthodologies innovantes d'utilisations des données de santé** et leurs limites, et sur les données de santé numériques.

Sommaire

Préface 3
18 recommandations pour favoriser l'utilisation des données de santé en vie réelle en recherche clinique 4
Contexte 9
Méthodologie de travail 12
Etat des lieux des acteurs, des utilisations et des pratiques 15
5 constats posés 23
Recommandations du Groupe "Données" (GT AIS/F-CRIN) 39
L'importance du "Modus operandi " - Mise en action des recommandations du Groupe "Données" (GT AIS/F-CRIN) 45

Contexte

Avec la loi de modernisation de notre système de santé (n° 2016-41) et son article 193 créant les conditions d'un accès ouvert aux données de santé et la loi (n° 2019-774) relative à l'organisation et à la transformation du système de santé qui a élargi le périmètre des données relevant du Système national des données de santé (SNDS), la France s'est dotée d'un arsenal législatif structurant. L'adoption en juin 2021 du plan « Innovation santé 2030 » dote la France d'une stratégie pour devenir la 1ère nation européenne innovante et souveraine en santé. Pour atteindre cet objectif, il est nécessaire d'agir à tous les stades de développement de l'innovation en santé, de la recherche fondamentale à la commercialisation et à la diffusion, en passant par la recherche translationnelle, la recherche clinique et la mise sur le marché.

Pour accélérer et optimiser le développement des produits de santé, l'Agence de l'Innovation en Santé (AIS) a identifié la recherche clinique comme levier prioritaire à mobiliser et s'est associée à l'Infrastructure nationale de recherche clinique, F-CRIN, dont l'un des axes prioritaires de sa nouvelle feuille de route « 2024/2028 » porte sur les essais cliniques pérennes.

Avec le lancement de travaux d'évaluation de « *l'évolution des méthodologies d'essais cliniques : nouveaux outils, nouveaux usages et conditions de recours* », il s'agit de **démontrer la valeur ajoutée de ces nouvelles approches à des moments clés du développement des produits de santé en vue de leur évaluation clinique. Ceci contribuera à créer les conditions d'acceptation de ces méthodologies de recherche clinique par les autorités réglementaires et agences d'évaluation des technologies de santé pour inciter les promoteurs d'essais à les utiliser.**

Ces méthodes innovantes pourraient permettre de résoudre des difficultés opérationnelles et de sortir d'impasses liées au schéma traditionnel de l'essai clinique randomisé prospectif comparatif mené par étapes successives. Si les processus d'évaluation des essais cliniques randomisés sont parfaitement standardisés et apportent un niveau de preuve très élevé, le fait que la recherche cible aujourd'hui des populations de plus en plus restreintes (passage du modèle du « blockbuster » au « custombuster ») complique le recrutement de patients en nombre suffisant, allonge les délais pour mesurer le critère principal de l'essai incompatible avec le besoin d'un accès rapide à l'innovation pour les patients ; ce qui a pour conséquence une complexité logistique et des coûts toujours plus élevés.

Les conditions et critères d'acceptabilité par les agences réglementaires dans le cadre des processus d'évaluation clinique sont dans certains cadres jugés peu favorables par les porteurs d'innovation en santé et les promoteurs de recherche. Elles ont déjà fait l'objet d'analyses mais pas encore de démonstrations pratiques, scientifiquement mesurées et quantifiées.

L'objectif du collectif réuni par l'AIS et F-CRIN est donc de prolonger les travaux théoriques antérieurs par l'évaluation comparée au travers du suivi de cas d'étude, de leur impact en matière de performance et de valeur avec des exigences similaires de preuve et de sécurité en précisant leurs conditions et modalités d'utilisation.

Les recommandations issues des réflexions du sous-groupe « Données » du GT AIS/F-CRIN ont vocation à nourrir la mise en place des objets financés dans le cadre de « France 2030 », notamment les cohortes, les Instituts Hospitalo-Universitaires (IHU) et les bioclusters.

Un axe essentiel de ces travaux concerne le niveau de qualité des données acquises en dehors d'essais cliniques, et plus particulièrement leurs modalités de structuration pour pouvoir les réutiliser et les mobiliser pour des projets de recherche innovants afin de contribuer au développement d'un produit de santé. Il est entendu que les données colligées dans le cadre d'essais cliniques ont également vocation à être réutilisées et rassemblées dans des bases de données, ce qui implique d'avoir réglé au préalable les questions de leur transfert et de leur standardisation. Ce dernier point fera l'objet de travaux ultérieurs.

Pour les données générées hors recherches cliniques, la vision du groupe s'est positionnée sur :

- ⇨ (1) la qualité et la gouvernance de ces données ; ces deux facteurs ayant un impact sur la pérennité et le potentiel d'utilisation des données des cohortes, bases de données, registres, entrepôts de données de santé
- ⇨ (2) la place du patient dans la gestion de ces données (information, absence d'opposition, consentement)
- ⇨ (3) la mise à disposition et l'accès élargi et facilité à ces données (fast-track pour l'accès).

Méthodologie de travail

Ce rapport repose sur une réflexion suivi du groupe de travail au travers d'une série d'entretiens menés auprès d'experts du domaine de la donnée, d'experts en santé, issus à la fois du public et du privé, et répartis comme suit (cf. paragraphe ci-après pour la liste des contributeurs auditionnés) :

Une grille d'audition a été établie par le groupe de travail autour des points suivants (cf guide d'audition en annexe) :

1. Les outils et technologies à mobiliser
2. La méthodologie de collecte et le niveau de qualité des données
3. L'utilisation des données
4. Les aspects réglementaires (information patient, CNIL…)
5. L'interopérabilité et le partage

Cette grille a servi de support d'échange semi-dirigé avec les structures auditionnées, l'ensemble des points n'ayant pas été traités par l'ensemble des structures.

Organisations auditionnées :

- **APHP** : Assistance Publique - Hôpitaux de Paris : le plus grand centre hospitalier universitaire d'Europe, acteur important de la recherche académique en France, jouant un rôle de producteur et d'utilisateur de données.
- **BNDMR** (Base Nationale de Données Maladies Rares) : centralise les données relatives aux maladies rares en France pour améliorer la prise en charge et la recherche.
- **CEGEDIM Health Data** : entreprise de technologie et de services, spécialisée dans la gestion des flux numérique, de l'écosystème santé et BtoB, ainsi que dans la conception de logiciels métiers destinés aux professionnels de santé et de l'assurance.
- **CNAM** : Caisse Nationale d'Assurance Maladie

- **Cohorte CONSTANCE** : grande cohorte nationale française sur la santé des adultes, axée sur l'étude des maladies chroniques et du vieillissement.

- **CUREETY** : start-up innovante qui développe des solutions numériques pour le suivi à distance, télésurveillance, des patients atteints de maladies chroniques.

- **France Cohorte** : initiative nationale visant à regrouper plusieurs cohortes de santé en France. Elle offre une source riche d'informations pour la recherche médicale.

- **Health Data Hub** (HDH) : plateforme nationale des données de santé, elle facilite l'accès aux données de santé pour la recherche, en respectant les règles de sécurité et de confidentialité.

- **HEVA** : entreprise spécialisée dans l'analyse des données de santé et notamment celles issues des bases médico-administratives comme le PMSI ou le SNIIRAM.

- **INCa** : Institut National du Cancer, pilote la lutte contre le cancer en France, coordonnant les actions en matière de recherche, de prévention et de soins.

- **INSERM** : Institut National de la Santé et de la Recherche Médicale, acteur clé de la recherche biomédicale en France

- **IQVIA** : multinationale américaine, fournisseur de solutions technologiques et de services autours des activités de recherche clinique et épidémiologique, spécialisée dans l'analyse des données de santé.

- **LIFEN** : Solution française spécialisée dans la gestion sécurisée des échanges de données médicales non structurées. Facilite la communication entre les systèmes d'information hospitaliers.

- **SKEZY** : entreprise spécialisée dans les solutions digitales pour le secteur médical, axée sur l'innovation et la facilitation des soins à distance.

- Plusieurs industriels en santé **(TAKEDA ; Janssen ; Amgen ; Merck ; BMS ; Pfizer)** : entreprises pharmaceutiques mondiales, développent des médicaments dans différentes pathologies pour lesquelles le besoin médical est élevé.

- **UNICANCER** : réseau des Centres de lutte contre le cancer (CLCC) dédié exclusivement à la cancérologie, promeut une approche multidisciplinaire de la recherche clinique axée sur le cancer.

Etat des lieux des acteurs, des utilisations et des pratiques

I. Les acteurs auditionnés

Au cours de la mission, une mosaïque d'acteurs répartis en six catégories décrites ci-dessous a été auditionnée, illustrant la pluralité de l'écosystème de production et d'utilisation des données de santé dans le champ de la recherche clinique.

Les **acteurs issus de l'industrie en santé** sont au nombre de trois. Leur activité se cristallise autour des études observationnelles ainsi que des essais cliniques de phase 1 à 4. Ces acteurs exploitent diverses sources de données, internes ou externes, qui sont autant de leviers pour le développement et l'évaluation des médicaments, constituant ainsi une étape cruciale dans l'innovation en santé.

Les **acteurs technologiques en données de santé** représentent quatre entités distinctes. Leur terrain d'innovation est la conception et la fourniture de technologies dédiées à la collecte, à l'analyse et à la gestion des données de santé. Ces acteurs déploient également des outils pour la télésurveillance et proposent des solutions logicielles de collecte qui sont des maillons essentiels dans l'écosystème des données de santé.

Pour la **recherche et innovation en cancérologie**, trois acteurs se distinguent. Ils œuvrent à élargir le spectre des connaissances sur les cancers, à catalyser l'innovation, et à accélérer l'accès aux avancées en cancérologie notamment en promouvant l'usage de l'intelligence artificielle.

La catégorie des **registres, cohortes et « banques de données »** est la plus représentée avec sept acteurs. Ils administrent des données essentiellement structurées de grande qualité utilisées dans les domaines de l'épidémiologie, la recherche clinique, et l'évaluation des parcours de soins. Ces structures alimentent de très nombreux projet de recherche et d'innovation en santé.

Les **établissements de santé auditionnés** comprennent deux groupements d'hôpitaux qui mettent en œuvre des entrepôts de données de Santé « EDS ». Ce sont des acteurs clés dans la collecte et l'utilisation des données de santé de vie réelle, contribuant ainsi à diverses finalités, notamment la recherche clinique et l'évaluation des pratiques médicales, la médecine personnalisée, les vigilances, le pilotage de l'offre de soin, et l'épidémiologie.

Enfin, les responsables du SNDS (CNAM ainsi que la Plateforme des Données de Santé (PDS) ou Health Data Hub), jouent un rôle clef à l'échelle nationale et européenne dans la coordination de l'écosystème pour la mise à disposition des données de santé françaises et la facilitation de leur utilisation pour des projets de recherche et d'innovation.

II. Usages des Données de Santé en Vie Réelle

Sources de Données

La pluralité des sources de données exploitées par les acteurs auditionnés illustre la richesse des informations accessibles dans le secteur de la santé. **Les registres, cohortes et bases de données médico-administratives** se positionnent comme des ressources-clefs d'informations pertinentes et de qualité sur des populations déterminées ou des pathologies spécifiques. Ces ressources sont principalement mobilisées pour des études épidémiologiques, l'évaluation des parcours de soin, et l'observance des traitements, avec un enrichissement notable grâce au chaînage avec des bases telles que la base principale du SNDS.

L'exploitation des **données issues des dossiers patients informatisés (DPI)** et plus généralement du **système d'information hospitalier** constitue une autre source d'information précieuse mais qui reste, en dehors de quelques initiatives en France, à l'état de « gisements ». Ces données, qui peuvent être collectées en quasi-temps réel au sein des établissements de santé dans les EDS conformes au référentiel de la CNIL ou autorisés par la CNIL[1], fournissent des données brutes. Elles sont considérées comme nécessitant une transformation et une mise en qualité préalable à leur exploitation ; leurs usages imposent des précautions méthodologiques adaptées aux objectifs, notamment en raison de leur niveau de qualité variable lorsqu'elles sont utilisées.

Ces données sont un levier essentiel pour l'évaluation des pratiques médicales mais aussi pour la recherche clinique. Le récent appel d'offre national pour la constitution des EDS et ceux fléchés pour développer leurs usages permettent à différents regroupements d'établissements de s'engager dans la réutilisation des gisements de données qu'ils produisent.

Les **données générées par les patients eux-mêmes**, à partir des objets connectés, des formulaires en ligne ou des applications mobiles, représentent une source d'information de plus en plus importante à prendre en compte. Les mesures PROMs (Patient-Reported Outcome Measures) et PREMs (Patient-Reported Experience Measures) permettent de capter l'expérience du patient, son ressenti, et d'autres paramètres de santé en conditions réelles, enrichissant ainsi la compréhension des maladies et des réponses aux traitements.,

Enfin, des données ouvertes et accessibles en open data sont également utilisées[2]. Ces données sont généralement des données agrégées et anonymisées. En France, plusieurs organismes publics mettent à disposition des données de santé en open data, c'est le cas notamment de l'Assurance Maladie, l'Inserm ou encore la Direction de la recherche, de l'évaluation et des études statistiques (DREES).

---

1 https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000044243098?init=true&page=1&query=%2A&searchField=ALL&tab_selection=cnil

2 https://www.snds.gouv.fr/SNDS/Open-Data

Diversité des usages

Les usages des données de santé en vie réelle par les acteurs auditionnés sont variés, allant de la recherche clinique aux études sur le coût des maladies, en passant par les études observationnelles, ou encore la conception de cohortes virtuelles. L'amélioration de la prise en charge des maladies chroniques à travers l'exploitation des données est un domaine d'intérêt majeur, tout comme le développement d'outils d'évaluation et d'aide à la décision aux échelles individuelles ou populationnelles. Les données permettent également d'analyser les attitudes et comportements des patients face à leur maladie, contribuant à apporter des informations sur les trajectoires des patients. Du fait de leurs richesses, les données issues d'entrepôts de données (EDS) hospitaliers portent actuellement l'éventail d'usages le plus large :

- Epidémiologie, analyse des soins courants, et veille sanitaire comme ce fut le cas pendant la pandémie de Covid-19,
- Recherche clinique
    - Faisabilité d'une étude et identification des centres de recherche
    - Optimisation des études : validation de critères de sélection, élaboration de critères de pré-sélection de patients (stratification)
    - Construction de bras contrôles externes
    - Etudes médico-économiques
    - Suivi de la performance des produits de santé émergents,
- Evaluation des pratiques et de l'offre de soins, vigilances, veille syndromique, médecine 5P, pilotage médico-économique

Tendances Notables

Certaines tendances se dessinent à travers l'usage des données de santé en vie réelle.

La **recherche clinique** est un domaine en expansion avec des modes opératoires méthodologiques en évolution, certains nécessitant la mobilisation des données disponibles ou à collecter. Les données personnelles des patients pourront être mobilisées pour la mise en place de bras contrôle externes, directement (émulation d'essai/ bras cibles, enrichissement des données, randomisation sur registre, …) ou être « transformées » afin de générer des données de santé artificielles (entrainement essai in silico, augmentation des cohortes par IA, …) tel que décrit dans le livre blanc « Données de santé artificielles : analyse et pistes de réflexion » publié en avril 2024.

Les données de vie réelle sont également indispensables pour l'entrainement des algorithmes d'intelligence artificielle permettant l'optimisation de la recherche clinique avec l'identification des patients répondeurs et la possibilité de stratification et l'élimination du bruit de fond dans les études cliniques. L'ensemble de ces méthodologies permettent in fine d'accélérer le développement des produits de santé tout en réduisant le coût.

De même, l'accent mis sur l'**interopérabilité**, le **chaînage de données multi-sources** et l'**échange des données** facilite la collaboration entre les différents acteurs et enrichit les possibilités d'analyses multicentriques et multidisciplinaires.

La diversité des sources et des usages témoigne de la dynamique d'innovation et de l'effort collectif pour exploiter au mieux les données de santé en vie réelle, en vue d'améliorer la connaissance et la prise en charge des pathologies. Cette dynamique s'inscrit dans un mouvement plus large visant à capitaliser sur la richesse et la complémentarité des données disponibles pour faire progresser la recherche et l'innovation en santé.

III. Technologies employées

Plateformes

La gestion des données de santé est centralisée au sein de différentes plateformes utilisées par les acteurs auditionnés. Quelques acteurs ont mis en place des architectures d'EDS ou de lacs de données (Data Lake), permettant ainsi une organisation et une centralisation des données. L'architecture EDS est particulièrement adaptée pour rassembler les données pseudonymisées, en association avec d'autres sources de données en open data. Les plateformes utilisées varient entre/ des environnements d'hébergement de données de santé (HDS), des hébergements sur cloud sécurisé, et des bulles sécurisées certifiées comme le HDH ou le CASD. Plusieurs acteurs mettent en œuvre des bulles homologuées appelées « système fils du SNDS » afin d'être plus autonomes dans les traitements et les appariements avec le SNDS, mais aussi avec les données de l'OpenData (ex. données de l'INSEE ou d'Airparif).

Il est à préciser que, même si au cours des auditions nous n'avons pas identifié d'expériences de plateformes fédérées ne nécessitant pas de centralisation des données, des initiatives prometteuses dans ce domaine ont récemment vu le jour, tel que le réseau DARWIN porté par l'EMA, permettant la réalisation d'études fédérées sur les bases de données des membres au préalable standardisées.

Outils et Logiciels de collecte et stockage de données

Les outils de **collecte de données** incluent des systèmes comme les eCRF (Electronic Case Report Form), qui sont encore très largement utilisés pour la collecte des données dans des études observationnelles ou complémentaires des essais de phase 3. Des logiciels spécifiques sont développés pour l'extraction de données à partir de logiciels de cabinets médicaux. D'autres outils technologiques ont été développés pour la collecte de données sensibles via une interface sécurisée et ergonomique, notamment des algorithmes de traitement automatique des langues (TAL). Ces outils

permettent par exemple la collecte de données de mesures physiologiques et plus largement d'informations clinico- biologiques.

La diversité des outils utilisés pour le **stockage des données** est pensée pour pouvoir couvrir l'ensemble des formats de données (structurées, semi-structurées et non structurées). Elle inclut des technologies de base de données SQL et NoSQL. Pour les architectures « de Lac », on voit que les technologies de traitement distribué du big data (ex. Hadoop, Spark) sont mises en œuvre essentiellement quand elles deviennent nécessaires, c'est-à-dire chez les acteurs devant analyser des données massives. Peu d'évolutions sont constatées dans les solutions utilisées pour le traitement et l'analyse, les entités auditionnées se reposant encore largement sur des outils classiques tels que SAS, R, RShiny ou des solutions propriétaires. On voit également se répandre l'usage d'outils et de méthodes OpenSource issues des data sciences. La volonté croissante de traiter des données hétérogènes et non structurées invite l'écosystème à s'équiper d'environnements adaptés aux techniques d'apprentissage automatique (machine learning). Ces environnements sont généralement composés de notebooks (tel que Jupyter), s'appuyant sur le langage Python et faisant appel aux librairies de cette branche de la science des données (Keras, Tensor Flow, SciKit-Learn, etc.). On remarque, par ailleurs, un intérêt de plus en plus marqué chez la plupart des acteurs pour le traitement du langage naturel pour tirer profit des données textuelles, qu'elles soient non structurées ou semi structurées, rendu possible avec ces nouvelles technologies.

Formats de données

Les formats de données standards utilisés pour l'interopérabilité et l'échange des données incluent des standards agnostiques (JSON, XML) ou spécifiques au domaine de la santé tels que OMOP, HL7 FHIR, PN13, DICOM, ou le format OSIRIS destiné au cancer et porté par un collectif interSIRIC. L'interopérabilité entre les formats des données de soins / données de vie réelle (ou real world data/RWD) et les formats des données d'études cliniques utilisés par les autorités d'évaluation est également un enjeu majeur ce qui implique la un travail de standardisation via par exemple le CDISC (Clinical Data Interchange Standards Consortium) par exemple.

On retrouve également des formats de données propriétaires adaptés à certains domaines spécifiques. Ces formats facilitent l'échange et l'interopérabilité des données entre différentes plateformes et acteurs et participent à la mise en qualité. Toutefois, certains formats laissent au producteur de données le libre choix des nomenclatures et de la codification des informations, pouvant induire un temps supplémentaire de contrôle des données et d'harmonisation des données par les analystes avant leur exploitation en cas de nomenclatures ou codification des données différentes.

La plupart des acteurs reconnaissent que la mise en œuvre pleine et entière de ces standards n'est pas encore effective. Même s'il suscite un intérêt marqué, le modèle OMOP n'a pas encore la maturité suffisante pour dépasser son cadre principal d'utilisation, celui des données centrées sur le patient pour les études observationnelles longitudinales, et couvrir l'ensemble des autres domaines et usages

des données de santé. Il fournit néanmoins, pour ce type d'études, des règles et outils pour faciliter et accélérer la transformation, le contrôle qualité et la réalisation des analyses spécifiques ou fédérées. Pour améliorer l'interopérabilité et la qualité de leurs données, plusieurs acteurs mentionnent avoir recours à des bases de connaissance, par exemple sur le médicament (ex : Base Claude Bernard, Thériaque) et l'utilisation des terminologies médicales standardisés (ex : CIM10, SNOMED CT, LOINC, MeSH) ou développées de façon ad hoc (notamment dans le champ des soins primaires).

Qualité des données

Même si certains travaux ont été menés au sein des acteurs auditionnés pour définir une démarche de mesure et de contrôle de la qualité de leurs données, il n'existe actuellement pas d'approche standardisée au niveau national, rendant difficiles les échanges entre acteurs à ce sujet, notamment entre utilisateurs et producteurs de données.

Certaines initiatives européennes ont été récemment lancées. On peut notamment citer un premier travail lancé par l'Agence Européenne du Médicament et visant à définir un modèle de qualité des données publié en fin 2022, un rapport publié en septembre 2023 issu du projet TEHDAS (« Toward the European health data space ») fournissant un ensemble de recommandations à ce sujet et, plus récemment, le lancement d'un projet par la commission européenne visant à créer un label sur la qualité et l'utilité des données de santé.

En synthèse, les technologies employées par les acteurs auditionnés démontrent une diversité dans l'approche de gestion et d'exploitation des données de santé. Les architectures technologiques mises en place, bien que variées, sont conçues pour répondre aux besoins de collecte, stockage, chaînage multi sources, mise en qualité, interopérabilité et de traitement des données, le tout dans un contexte de sécurité adapté. Des méthodes d'intelligence artificielle à base de règles ou de machine learning sont peu à peu introduites pour la mise en qualité, l'analyse des données ou la modélisation des données, ouvrant la voie à de nouvelles sources (données non structurées) et de nouveaux usages. Les questions de la standardisation et de la mesure de la qualité des données nécessitent cependant des réflexions plus globales, permettant un meilleur alignement des pratiques et une plus forte interopérabilité des résultats produits.

Au terme de cet état des lieux, cinq constats peuvent être posés :

1. **La force d'un écosystème de données de santé riche et varié**
2. **Des procédures de collecte, de stockage et mise en qualité rigoureuses**
3. **Des contraintes réglementaires partiellement gérées par le développement d'outils innovants**
4. **Une utilisation des RWD (Real World Data) largement valorisée**
5. **Des freins à l'utilisation des données de santé en recherche clinique**

5 constats posés

1er constat : La force d'un écosystème de données de santé riche et varié

La France dispose d'un écosystème de données de santé diversifié, porté et représenté par des acteurs multiples. Cet écosystème s'est renforcé en 2016 avec la mise en place du Système national des données de santé (SNDS) composé des données de remboursement des régimes d'Assurance maladie obligatoires (SNIIRAM) mais également d'autres données comme les données des hôpitaux dans les bases du programme de médicalisation des systèmes d'information – PMSI ou les des bases de données sur les causes médicales de décès (base du centre d'épidémiologie sur les causes médicales de décès de l'institut national de santé et de la recherche médicale – CépiDc de l'Inserm.

La loi du 24 juillet 2019 a ajouté à ce « SNDS historique » de nouvelles catégories de données, telles que les données issues de la prise en charge médicale lorsque les actes sont remboursés par la sécurité sociale, les données issues des visites médicales scolaires, des services de protection maternelle ou infantile, des visites de santé au travail ou des enquêtes appariées, etc. L'ensemble de ces données entre dans le périmètre du « SNDS élargi ». Lorsqu'elles sont utilisées notamment à des fins de recherche, elles sont soumises au respect de certaines dispositions du code de la santé publique, tout particulièrement au respect du référentiel de sécurité, à l'interdiction de poursuite de finalités interdites, aux conditions d'accès aux données via un accès permanent ou à la réalisation d'une formalité auprès de la CNIL ainsi qu'aux modalités de transparence.

Ce rattachement juridique au « SNDS élargi » n'emporte pas pour autant à lui seul de migration des données au sein d'une base centralisée. Seules certaines données du « SNDS élargi » sont centralisées par la CNAM et/ou la PDS :

- La base principale du SNDS composée des données du « SNDS historique » (enrichi avec les données relatives à la Covid-19 extraites des bases Vaccin Covid et SI-DEP (système d'information et de dépistage) ;
- La base catalogue composée de dix bases de données prévues par un arrêté du 12 mai 2022.

Reprenant la proposition du rapport Villani et le rapport Polton/Begaud, la loi de 2019 crée la Plateforme des données de santé (ou Health Data Hub). Ce dernier, a pour rôle de contribuer à fluidifier l'accès aux données de santé pour permettre l'utilisation de ces dernières à des fins de recherche et notamment accompagner le développement de l'IA sur des données partagées. L'accès aux données de la base principale du SNDS, au même titre que l'accès aux données collectées dans le cadre du soin, est un réel enjeu stratégique, tant les données de santé sont devenues des facteurs d'attractivité à l'échelle internationale, des vecteurs de progrès scientifique et d'enrichissement des connaissances médicales et scientifiques. Ces données de

santé sont à l'heure actuelle utilisées, et ce de manière croissante, afin mieux comprendre l'histoire naturelle de la maladie, de recueillir des données épidémiologiques, de mieux cibler les besoins et ainsi d'optimiser la conception des plans de développement, de fournir des groupes comparateurs externes ou encore appréhender la valeur d'une thérapie en « vraie vie ». L'utilisation de « Real world data » dans la démarche de recherche clinique permet (1) d'apporter une meilleure compréhension des soins standards susceptibles d'être utile au choix et/ou à la construction d'un bras de référence d'un essai clinique randomisé ou également (2) de mieux maitriser et pré-identifier les populations potentiellement accessibles au recrutement dans un essai clinique participant à l'étude de faisabilité d'un essai clinique.

Ces enjeux ont été confirmés par le discours du Président de la République lors de la présentation d'Innovation Santé 2030 qui mettait les données au cœur de sa stratégie.

Ces améliorations ont été accompagnées d'un changement dans la manière d'exploiter ces données, évoluant ainsi d'analyse base par base à la mise en commun de données de santé émanant de plusieurs systèmes d'information médicaux, notamment à travers la création d'entrepôts de données de santé (EDS).

Cet écosystème, au fil des évolutions législatives, s'est également enrichi d'une multitude d'acteurs publics et privés, travaillant de façon complémentaire à la collecte, au stockage, à la mise en qualité et à l'exploitation et l'utilisation de ces données de santé.

2ème constat : Des procédures de collecte, de stockage et mise en qualité rigoureuses

Les acteurs de l'écosystème précédemment décrit ont le plus souvent à disposition un grand nombre d'outils et de technologies, parfois développés par eux-mêmes, pour collecter, stocker, traiter et analyser les données de santé.

Un certain nombre d'acteurs privés a, à ce titre, créé des solutions à destination d'acteurs (publics ou privés) afin de faciliter et d'optimiser ces activités, très consommatrices de temps humain. Les procédures suivies par les acteurs auditionnés se sont toutes distinguées par leur rigueur, leur précision mais aussi leur complexité. En effet, la qualité de la collecte au même titre que la qualité de la donnée elle-même, est un des principaux garants de l'excellence et la pertinence de la recherche scientifique qui en découle. Par ailleurs, les données de santé étant par nature de données qualifiées de sensibles, il est primordial de mettre tout en œuvre pour garantir la protection des données personnelles.

Ces phases de contrôle sont élaborées en une succession d'étapes distinctes et standardisées utilisant des indicateurs et des référentiels parfois communs mais souvent propres à chaque acteur. Ces étapes, bien qu'indispensables, sont complexes et chronophages, et nécessitent encore à l'heure actuelle des ressources humaines importantes, le plus souvent très qualifiées. Fort heureusement, l'émergence des

solutions automatisées, de bibliothèques d'algorithmes, de modèles d'apprentissage associés à des capacités de calculs importantes, permettent d'accélérer et d'optimiser ces étapes tout en conservant un contrôle humain justifié mais ciblé. Leur usage est à favoriser.

Le niveau de qualité de la donnée est étroitement lié à l'accessibilité et à la lisibilité des différentes procédures qui ont conduit à son stockage et son exploitation. La disponibilité d'une documentation accessible et la connaissance précise des modes et conditions de collecte des données sont ainsi indispensables et reste encore un enjeu d'actualité. De la même manière, un retour vers les « producteurs » ou les « fournisseurs » de données est un gage de qualité et de cohérence, seuls acteurs à même de parfaitement maîtriser la qualité et les faiblesses des données partagées. Il est important de garder en tête la nécessité de définir, dans la majeure partie des cas, la question et l'objectif avant de collecter des données de santé, et non l'inverse.

Quelques modèles de format type OMOP ou eHOP se dégagent mais aucun modèle de données standard n'est retrouvé de manière uniforme dans les structures auditionnées. Cependant, cette diversité de format ne semble pas freiner l'élan d'exploitation et de recherche scientifique, étant toujours possible de proposer si requis une adaptation (par matching) même si complexe d'un format à un autre.

Bien que la donnée structurée soit celle le plus souvent collectée, une **montée en puissance de l'utilisation de la donnée non structurée**, beaucoup plus riche par nature, s'entrevoit chez un grand nombre des acteurs auditionnés. Afin d'avoir la capacité d'exploiter ces données non structurées, il est indispensable de faire appel à nouveau à des ressources très qualifiées, des mathématiciens ou bio-informaticiens spécialisés en intelligence artificielle et technique de traitement automatique du langage naturel en lien avec un expertise médicale continue.

3ème constat : Des contraintes réglementaires partiellement gérées par le développement d'outils innovants

Malgré des modalités juridiques et techniques d'accès aux données de santé jugées complexes et des délais trop longs, les acteurs du domaine ont déployé des efforts importants pour leur permettre de réutiliser les données issues du soin.

I. Le régime juridique applicable à la réutilisation des données

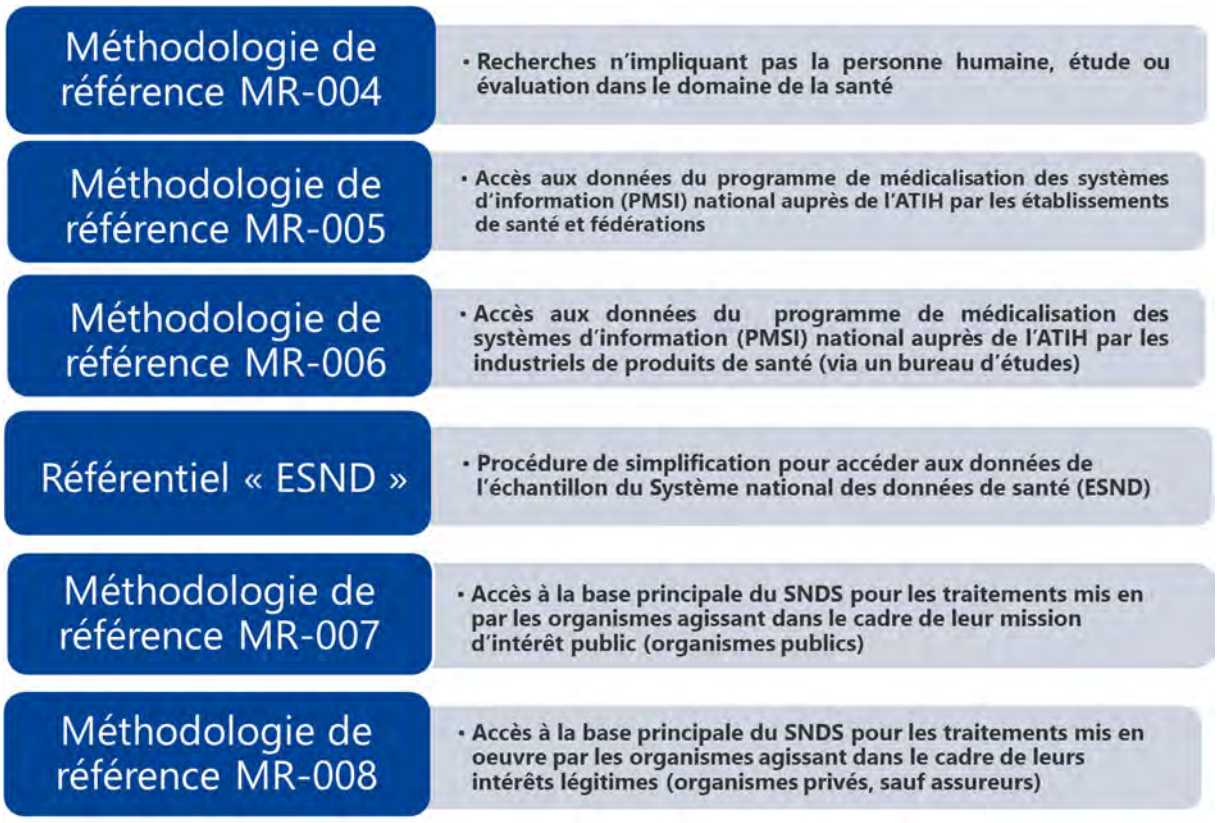
La simple réutilisation de données déjà collectées (dans le cadre du soin ou d'une précédente recherche) ne nécessite pas le recueil d'un consentement des personnes concernées pour mettre en œuvre une recherche n'impliquant pas la personne humaine. Dans cette hypothèse, les personnes doivent être informées de cette réutilisation et pouvoir être en mesure de s'y opposer.

Il est à noter que le cadre juridique est toutefois différent pour les essais hybrides impliquant une réutilisation des données des patients dans le bras contrôle ou témoin, cette réutilisation obéissant alors au même régime juridique que celui applicable au bras principal (et notamment à la nécessité de recueillir un consentement en vue de la participation à l'étude lorsque la recherche est interventionnelle, conformément aux dispositions du code de la santé publique) ; ce qui n'est pas sans poser de difficultés pratiques.

II. Les formalités à réaliser auprès de la CNIL

Qu'elles impliquent ou non la personne humaine, les recherches scientifiques en santé nécessitant un traitement de données de santé, doivent, sauf exceptions prévues par la loi, faire l'objet d'une déclaration de conformité à une méthodologie standard de traitement de données décrite dans un référentiel publié par la CNIL appelé « méthodologie de référence ». A date, neuf référentiels sont applicables aux recherches en santé pour encadrer la plupart des traitements. Par exception, les recherches qui ne sont pas conformes à ces référentiels en raison d'une sensibilité particulière (traitements de donnés d'infractions par exemple) doivent être autorisées par la CNIL, après avis du comité compétent (comité de protection des personnes ou Comité éthique et scientifique pour les recherches, les études et les évaluations dans le domaine de la santé).

Parmi ces neuf référentiels, six d'entre eux concernent les recherches nécessitant de réutiliser des données déjà collectées :



Source CNIL

En particulier, et afin de faciliter les réutilisations de données, la méthodologie de référence MR-004 prévoit que des données issues du soin ou d'une précédente étude puissent faire l'objet d'une réutilisation sans qu'il soit procédé à une nouvelle information individuelle des personnes. Il en va ainsi notamment lorsque l'information délivrée lors de la collecte initiale des données prévoit la possibilité de réutiliser les données et renvoie à un dispositif spécifique d'information auquel les personnes concernées pourront se reporter préalablement à la mise en œuvre de chaque nouveau traitement de données (par exemple : un site web ou une page dédiée appelé « portail de transparence » sur lequel serait présenté chaque projet de recherche mené sur les données collectées dans le cadre de l'information initiale). La MR-004 n'impose pas que le champ thérapeutique des études ultérieures soit précisé dans la note d'information initiale.

Cette méthode nécessite néanmoins d'avoir anticipé la possibilité de réutilisation dès la conception du traitement initial en créant une page « web » dédiée à l'information. La mise en place de ces portails de référence permet de s'affranchir de l'envoi de nouvelles notes d'information pour chaque réutilisation de données.

Des portails spécifiques ou pages web dédiés aux patients sont fréquemment construits permettant ainsi de partager en continu des éléments quant à la réutilisation des données de santé.

De plus, il faut avoir la capacité d'anticiper l'accès simultané à différentes sources de données pour répondre à des objectifs pour lesquels les patients n'ont pas été informés du caractère « multisource » lors de l'information initiale. Ces approches transversales à partir de plusieurs sources de données de patients, ayant eu un niveau d'information adapté par source de données, restent complexes et correspondent à des besoins nouveaux. Par exemple des travaux s'appuyant sur les données d'un EDS avec des données d'essais cliniques, des données de soins complémentaires ou issues d'une autre source de données, impliquent de prendre un compte un niveau adapté d'information des patients, si possible en amont sinon a posteriori.

En cas de réutilisation de données d'un même patient provenant de plusieurs sources de données, la question de l'articulation et de l'alimentation de ces différents portails de transparence pourrait se poser.

D'autres acteurs tels les associations de patients prennent une place de plus en plus importante dans cet écosystème des données de santé. Elles sont notamment sollicitées dans ces procédures d'information et de communication auprès des patients. Cela est d'autant plus vrai dans le domaine des maladies rares, précurseur dans ce domaine. L'implication de représentants de patients devrait fortement se démocratiser dans les années à venir, cela ne venant que renforcer le caractère transparent et éthique des démarches de réutilisation des données de santé.

4ème constat : une utilisation des RWD (Real World Data) largement valorisée

I. Données structurées/non structurées

A l'heure actuelle, si la majorité des données exploitées sont des données structurées, la part de données non structurées augmente progressivement (données textuelles, données numériques d'imagerie, lames pathologiques, …).

Les études qui en découlent se positionnent dans la majeure partie des cas soit en suivi post commercialisation pour connaître l'usage en vie réelle d'une nouvelle thérapie et ainsi appréhender la valeur de cette thérapie, soit en phase de développement clinique afin de caractériser une pathologie (étude à visée épidémiologique) et sa prise en charge, voire son fardeau clinique ou économique.

Il existe ainsi une multiplicité des activités scientifiques et des utilisations de ces données allant désormais jusqu'à la combinaison de plusieurs sources de données entre elles.

Un chaînage entre des registres, cohortes et la base principale du SNDS est aussi de plus en plus fréquent malgré les délais d'accès aux données qui s'allongent. Ces chaînages peuvent être réalisés de façon probabiliste ou déterministe, via le numéro de sécurité sociale des patients ; le plus souvent en suivant les circuits types publiés par la CNIL.

Outre les appariements « ponctuels » pour des projets de recherche, il est également possible, de chaîner de façon pérenne des données d'entrepôts de données de santé ou de registres avec celles du SNDS : « Cardiohub » ; « Magellan » ; plateforme AGORIA ; « Data-mesh » ; « Colibri-Pneumo », « UroCCR-Chain », plateforme de cancérologie de l'INCa, « Harmonie » d'Unicancer… .Il faut bien anticiper lors de la collecte de données et de la constitution des EDS la capacité technique et réglementaire à effectuer ce chainage nécessaire pour de nombreux projets.

En pratique, une fois les données déjà chaînées, elles peuvent être réutilisées (notamment dans le cadre de la MR-004 – si toutes les autres conditions, notamment en termes d'information – sont remplies). De nombreux EDS chaînés aux données du SNDS ont fait le choix de délivrer une information individuelle aux personnes concernées et de les renvoyer vers un portail de transparence afin d'être informées des usages secondaires afin de pouvoir mobiliser la MR-004 pour les études ultérieures.

Les données collectées dans le cadre d'un accès précoce, là aussi spécificité très française, pourraient être une source précieuse d'information, à condition qu'elles soient exhaustives et de qualité, venant enrichir les dossiers soumis aux autorités en vue d'un remboursement et intéressent fortement nos voisins européens. A l'issue d'une collecte de données dans un cadre d'accès précoce, la poursuite du suivi des patients dans une phase « post accès précoce », avec parfois des données

complémentaires, doit être favorisée et identifiée comme une source de données contributive.

L'émergence d'une utilisation plus précoce dans les phases de développement clinique s'accompagne de la mise œuvre de méthodologies et d'approches innovantes telles que les bras synthétiques ou les essais émulés. A titre d'exemple, parmi les 34 essais cliniques in-silico réalisés en 2023 dans le domaine de la cardiologie, 6 évaluaient un médicament[3].

L'exploitation de ces données en vie réelle se valorise de mieux en mieux au plan scientifique avec des communications écrites ou orales en congrès, à l'échelle nationale ou internationale et des publications d'articles scientifiques dans des journaux avec comité de relecture international de haut niveau. Par ailleurs, ces données sont largement valorisées en interne au sein d'organisations de taille importante où l'accès à ces données et la conduite d'études sont des vecteurs d'attractivité et de compétitivité. Par ailleurs les analyses de ces données, même si elles ne conduisent pas systématiquement à des publications, peuvent influencer un grand nombre d'actions ou piliers stratégiques en leur sein.

Les autorités de santé (réglementaires et/ou de remboursement) locales ou européennes commencent elles aussi à intégrer progressivement dans leurs doctrines la prise en compte des données en vie réelle :

- La Haute Autorité de Santé (HAS) définit les données de vie réelle comme issues d'études observationnelles, bien que ce soit une définition très restrictive au regard du cadre plus global des données de vie réelle qui peuvent avoir différentes sources (autorisation d'accès précoce/AAP ; autorisation d'accès compassionnel/AAC ; registres, requêtes de bases de données telles que le Système national des données de santé ; études post-inscription ; etc.).
- Au niveau européen, l'EMA prend en compte depuis plusieurs années les RWD, que ce soit dans les dossiers de demande d'AMM ou les dossiers de demande d'extension d'indication. Des études menées sur des RWD ont été considérées comme contribuant au niveau de preuve sur un niveau d'efficacité pour la décision[4] ; leur nombre ne pouvant qu'augmenter dans les années à venir.

Pour les études post-inscription, bien qu'elles soient rarement adaptées en termes de niveau de preuve pour la revendication d'une ASMR supérieure à celle obtenue initialement sur la base de données expérimentales, ces données peuvent contribuer à une valorisation ou au maintien du SMR ou de l'ASMR et apportent généralement des informations cruciales pour recommander les modalités de bon usage des médicaments. Les développements méthodologiques actuels spécifiques aux études sur données de vie réelle permettent de minimiser de nombreux biais et d'optimiser la validité interne de ces travaux, sachant la force de leur validité externe.

[3] A systematic review of cardiac in-silico clinical trials - PubMed (nih.gov)
[4] (Bakker E, ASCPT, 2022 ; https://doi.org/10.1002/cpt.2766)

Dans la valorisation de l'ASMR, les données issues d'études en vie réelle peuvent être prises en compte au même titre qu'un contrôle externe dans les essais cliniques mono-bras si ces comparaisons sont bien entendu de bonne qualité méthodologique.

II. Une nécessaire collaboration et approche partenariale

Un aspect important à mentionner est la notion de **collaboration** qui caractérise le plus souvent l'environnement des données en vie réelle, contrairement à celui des essais cliniques portés par un seul promoteur. Cette collaboration, nécessaire, met en relation des acteurs complémentaires, le plus souvent publics et privés, qui œuvrent ensemble pour répondre à des questions scientifiques précises en déployant souvent des approches innovantes pour faciliter et garantir un accès aux données et en extraire le maximum d'informations pertinentes. Cette approche partenariale est souvent l'occasion d'intégrer les associations de patients qui jouent un rôle important pour faciliter la collecte de données et optimiser la compréhension auprès des patients sur l'intérêt majeur de conduire ces études.

Les données de santé sont un trésor pour la science. Leur exploitation de plus en plus variée qui intègre notamment les données non structurées (représentant à peu près 80% des données hospitalières) nécessite l'étroite participation et collaboration d'acteurs divers (structures académiques, firmes privés, start-ups spécialisées en IA, associations de patients etc.) maitrisant une ou plusieurs catégories de données (EDS, registre, SNDS, PROM, données numériques…). Les données de santé et leur accès sont, au-delà d'un enjeu d'attractivité et de souveraineté, une absolue priorité pour la communauté. Il est ainsi crucial que la puissance publique mette les moyens pour continuer d'enrichir cet écosystème mais surtout de faciliter son accessibilité au plus grand nombre, tout en continuant de garantir la protection des données personnelles.

5ème constat : des freins à lever à l'utilisation des données de santé en recherche clinique

La ré exploitation des données générées lors des activités de soins, RWD, est un élément clé pour l'avenir de la recherche en France et au niveau européen.

Elle s'intègre dans la démarche d'évaluation du bénéfice/risque en recherche clinique pour fournir des éléments de preuve pour une décision éclairée. L'objectif de personnalisation des traitements pour optimiser la prise en charge de chaque patient reste un objectif clef à faire progresser.

Celui-ci doit s'appuyer sur l'ensemble des données clinico- biologiques, médico-administratives, numériques et sur toute donnée générée par les patients en lien avec leur prise en charge. Ces données doivent permettre de clairement intégrer les éléments de la prise en charge hospitalière comme de la médecine de ville à partir de l'ensemble des sources de données disponibles, qu'elles soient académiques ou

privées. Aujourd'hui, le cadre réglementaire, la maitrise des technologies de stockage et d'analyse, la capacité à rapprocher les données de typologies différentes (centralisation et/ou interopérabilité), en parallèle d'une acculturation des acteurs dont la puissance publique, permet de générer des informations scientifiques utiles à la décision en santé. Ces données doivent trouver un juste niveau de reconnaissance et d'utilisation pour participer à l'amélioration de l'accès à l'innovation, tenant compte de leurs forces mais aussi de leurs contraintes.

Sont ici abordés les obstacles identifiés par les différents acteurs de la donnée en s'appuyant sur des points concrets et freins identifiés notamment dans son utilisation en lien avec la recherche clinique.

Une enquête réalisée en 2023 par le Leem, syndicat de l'industrie du médicament, auprès de ses adhérents dans le cadre d'une étude sur : « L'attractivité de la France pour l'usage secondaire des données de santé » a ainsi confirmé le fort intérêt des industriels du médicament pour les données médicales. Elle a toutefois mis en lumière les délais de mise à disposition en France jugés trop longs en particulier pour les données issues du SNDS (seules 25% des demandes sont satisfaites dans un délai d'1 an) ainsi qu'un taux important de renonciation d'utilisation : 2/3 des industriels interrogés déclarent avoir dû renoncer à posteriori à tout ou partie de leurs projets de recherche). Les raisons de cette renonciation sont à rechercher dans les délais d'accès mais aussi à une "déception" à réception des sets de données sur leur utilité/utilisabilité dans le cadre de l'étude.

I. Une croissance exponentielle des données et des freins techniques

Les capacités techniques à rapprocher les données et à mener des analyses sur des données de typologies différentes dépendent des conditions de gestion de leur stockage et de maintien d'accès maitrisés. En effet les différentes typologies de données nécessitent un format de communication entre ces données, une interopérabilité, et si possible un format commun. A ce jour les formats portés par l'UE, dont OMOP, sont complexes à mettre en œuvre et à maintenir sachant qu'il y existe en amont de chaque base ou entrepôt de données de santé (EDS) presque toujours un format commun ou propriétaire adapté en priorité aux besoins d'exploitation du producteur/agrégateur de données. Les référentiels utilisés en France (CCAM, CIM-10, CIM-O…ou référentiels propriétaires) ne sont pas utilisés de façon homogène pour permettre une interopérabilité simple des données. Cela est vrai pour les données au niveau national et plus encore entre différents pays.

Le volume de ces données est également un sujet de discussion. Aujourd'hui la capacité de stockage se compte de plus en plus en téraoctets voir en pétaoctets pour tenir compte des données numérisées en croissance constante (imageries, lames pathologiques, données de génomique), ces données devant également communiquer avec des données thérapeutiques et clinico-biologiques mieux maitrisées.

La centralisation ou la mise en communication des données par des processus d'interopérabilité entre les sources de données reposent sur un hébergement centralisé ou fédéré. Celui-ci doit être capable d'apporter les services nécessaires permettant d'accéder, de partager et d'analyser des données en nombre croissant de manière exponentielle. Les 2 modes d'hébergement, hébergeur de données de santé « on-premise » ou Cloud, ont des référentiels de sécurité exigeants et nécessaires. Par exemple, la norme SecNumCloud en lien avec le contexte de souveraineté des données implique un investissement important des hébergeurs qui développent actuellement leurs offres, impliquant également une composante d'infogérance complexe et une maitrise parfaite des droits des utilisateurs. En cas de traitement de données du SNDS, un référentiel de sécurité ad hoc doit également être respecté.

II. L'identification insuffisante des sources de données et des outils de recherche

L'identification des sources de données est un préalable important pour tout chercheur et promoteur. L'absence de référencement systématique et correctement documentée des différentes sources de données potentiellement accessibles, (qui ont leurs propres règles d'accès, de gouvernance et de modèle financier), reste un frein pour de nombreux promoteurs d'étude. A titre d'exemple, du fait de lacunes dans la compréhension de l'organisation et du potentiel des données de la base principale du SNDS, celle-ci reste sous-exploitée en épidémiologie descriptive mais également en recherche clinique où elle pourrait être une source de données pertinentes.

La documentation complète de chaque base/EDS n'est pas toujours simplement accessible. La conception d'études spécifiques en lien avec des données existantes se développe, que ce soit pour des études avec émulation de bras contrôle a priori ou a posteriori ou d'études spécifiques sur RWD. Sans un référencement informatif accessible avec un lien direct au producteur/agrégateur de données, le chercheur/promoteur n'a pas toujours les connaissances sur le potentiel de données accessibles pour répondre aux objectifs de son projet de recherche. Ceci explique que des méthodologies et conceptions d'études utilisant notamment la reconstruction de bras contrôle à partir de données issues de bases et EDS, et des techniques d'analyses statistiques propres aux analyses sur RWD (scores de propension, variables instrumentales, méthodes de contrôles négatifs,..) ne soient pas toujours mises en œuvre et prises en compte, par les chercheurs/promoteurs ou les destinataires des recherches (relecteurs des revues de haut niveau ou autorités de santé).

Le développement des modèles faisant appel à l'intelligence artificielle (IA) entrainée sur des RWD, offre également une opportunité pour accélérer l'accès à l'innovation dans un nombre de domaines de plus en plus important. Cependant, en l'absence d'identification, de classification, de référencement de la validité des IA, celles-ci restent encore au stade de la recherche dans un cadre expérimental. Cela devrait rapidement évoluer si les sujets éthiques et de connaissance fine de chaque IA gagnent en transparence.

Bien que les bases de données et EDS constituent un apport majeur pour de nombreux projets de recherche, une mauvaise maîtrise des sources de données et de méthodes de conception et d'analyse statistique freine l'innovation en santé.

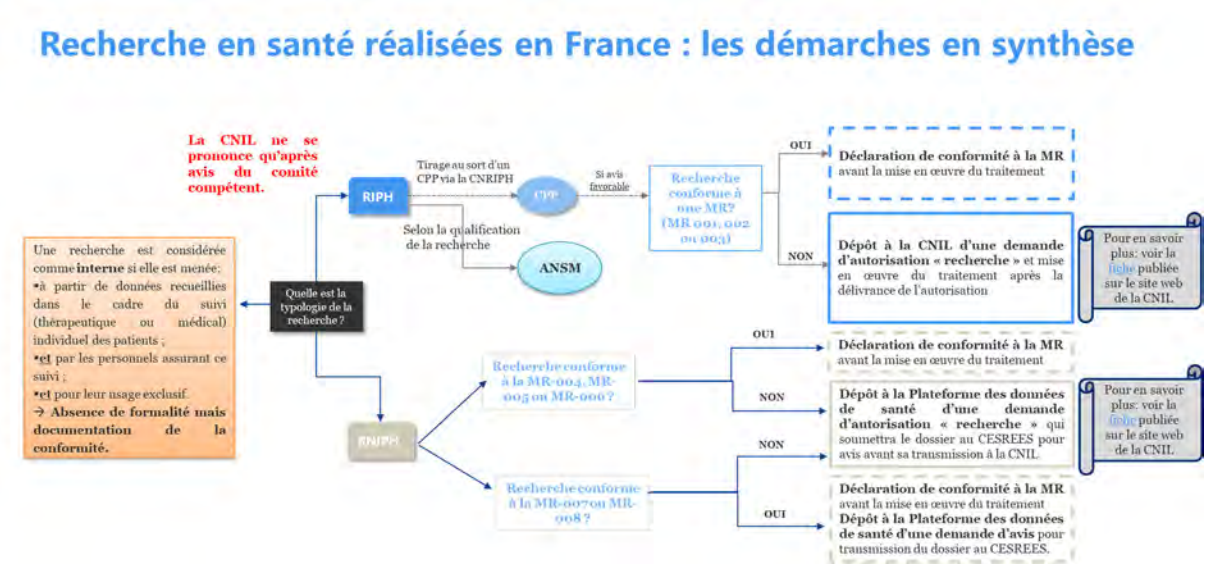
III. Les problématiques institutionnelles et éthiques

Des acteurs et des circuits de demande d'accès identifiés : le Health Data Hub, malgré les discussions sur le choix actuel d'hébergement des données joue un rôle important comme guichet unique pour la gestion et l'accompagnement des demandes d'accès aux données RWD, notamment à la base principale du SNDS.

Dans le cas des recherches n'impliquant pas la personne humaine, Le CESREES est chargé de se prononcer sur la finalité et la méthodologie de la recherche, la nécessité du recours à des données de santé à caractère personnel, la pertinence éthique, la qualité scientifique du projet et, le cas échéant, sur le caractère d'intérêt public que présente le projet dans deux cas :

- lorsque l'étude n'est pas conforme à un référentiel de la CNIL (MR-004, MR-005, MR-006, MR-007, MR-008 ou au référentiel « ESND »),
- lorsque l'étude est conforme aux MR-007 et MR-008 applicable aux traitements impliquant uniquement une réutilisation de données issues de la base principale du SNDS.

La CNIL homologue des référentiels (dont les méthodologies de référence) et autorise les traitements qui ne sont pas conformes à ces référentiels. Dans l'hypothèse où de nouveaux designs d'essais, qui se standardiseraient, impliqueraient la mise en œuvre de traitements non conformes à une MR de la CNIL, il est proposé d'en intégrer certaines spécificités dans les MR en vue d'éviter un régime d'autorisation systématique.



Source : CNIL

Le processus de gestion des demandes d'accès, que ce soit au SNDS, à un registre ou à un EDS suit donc un **circuit parfaitement identifié**, bien que les producteurs/agrégateurs de données qui sont seuls à avoir les connaissances fines sur « leur » base/EDS pourraient certainement en étant plus impliqués dans ce circuit, optimiser encore les délais et les demandes de projets. En effet de multiples allers-retours liés à une méconnaissances des données et de leur potentiel peuvent retarder ou même interrompre des études, ralentissant ainsi les avancées scientifiques. Ces délais peuvent également engendrer des coûts additionnels, et parfois même décourager des initiatives innovantes. Ces lenteurs, outre leur impact financier, contribuent à éroder la confiance des patients et du public qui espèrent des avancées médicales d'autant plus si les données permettant ces avancées de connaissances sont accessibles.

Disparité de connaissances au sein des acteurs impliqués dans le circuit de décision de mise en place de ces recherches : il est parfois difficile d'avoir une approche différenciée entre un projet mené sur données issues de protocoles d'essais cliniques parfaitement contrôlés avec données protocolaires, monitorées et exhaustives, d'un projet sur des « RWD » où les données sont quantitativement plus importantes, riches, mais non homogènes ou moins qualifiées à leur source. Cet écueil est également observé quand on s'intéresse à des projets impliquant des métadonnées, ou données générées par des intelligences artificielles. Les « RWD » apportent des informations avec leur propre niveau de validité, utiles à la collectivité lorsqu'elles sont exploitées dans leurs cadres, avec leurs limites, pour aider à la décision et permettre d'enrichir les connaissances scientifiques de haut niveau. L'acculturation et la nécessité de mise au niveau de l'ensemble des acteurs et pour commencer, en priorité, les personnels médicaux et paramédicaux, restent encore à ce jour un frein qui doit progressivement être levé.

L'accès aux données doit s'appuyer sur une gouvernance cohérente et acceptée par l'ensemble des acteurs impliqués, tout en tenant compte du cadre règlementaire sur leur réutilisation. L'équilibre entre le nécessaire partage des données indispensables pour les chercheurs/promoteurs et l'importance de la protection des données des patients repose sur un équilibre à mieux définir.

IV. Les enjeux réglementaires et la place du patient :

Il convient de distinguer les **règlementations sectorielles applicables aux recherches** (code de la santé publique pour les RIPH et les traitements de données du SNDS, règlements relatifs aux essais cliniques et aux dispositifs médicaux, y compris de diagnostic in vitro…) **de la règlementation spécifique applicable en matière de traitements de données à caractère personnel** (RGPD et loi du 6 janvier 1978 modifiée dite « informatique et libertés »). Selon le contexte de la collecte initiale des données, plusieurs règlementations peuvent avoir vocation à s'appliquer : ainsi, les dispositions de l'article L. 1122-1 code de la santé publique prévoient la liste des informations qui doivent être transmises aux participants afin que leur consentement

en vue de leur participation à la recherche, lorsqu'il est requis, soit suffisamment éclairé.

S'agissant des traitements de données mis en œuvre à des fins de recherches, études ou évaluations en santé, l'article 69 de la loi « Informatique et libertés » prévoit, par principe, la délivrance d'une information individuelle des personnes concernées conformément au RGPD.

Les articles 12 et suivants du RGPD imposent de fournir aux patients une information claire et transparente sur les usages de leurs données afin d'assurer une loyauté des traitements. A ce titre, un certain nombre d'informations doivent être transmises aux personnes concernées, que les données aient été collectées directement (article 13 du RGPD) ou indirectement (article 14 du RGPD) auprès d'elles. Il en va ainsi notamment :

- De l'identité ainsi que les coordonnées du responsable du traitement et du délégué à la protection des données ;
- Des finalités (objectifs) du traitement et sa base (fondement) juridique ;
- Des personnes ayant accès aux données (destinataires) ;
- De l'existence d'un transfert de données personnelles en dehors de l'Union européenne ainsi que ses modalités d'encadrement ;
- De la durée de conservation des données ;
- Des droits dont elle dispose (accès, rectification, effacement, limitation, opposition et portabilité des données) ;
- Le droit d'introduire une réclamation auprès de la CNIL.

Par ailleurs, dans le cas d'une prise de décision automatisée, les personnes doivent être informées, lors de la collecte de leurs données et à tout moment sur leur demande, de son existence, de la logique sous-jacente ainsi que de l'importance et des conséquences prévues de cette décision.

Si les données ne sont pas recueillies directement auprès des personnes concernées, leur origine (source) doit être mentionnée dans la note d'information.

Lors de la collecte initiale des données dans le cadre d'une recherche impliquant la personne humaine, les dispositions de l'article L. 1122-1-2 du code de la santé publique prévoient que le promoteur peut demander au participant au moment où celui-ci donne son consentement éclairé (lorsqu'il est requis) d'accepter que ses données soient utilisées lors de recherches ultérieures exclusivement à des fins scientifiques. La personne peut retirer son consentement à cette utilisation ultérieure ou exercer sa faculté d'opposition à tout moment.

Lorsque les thématiques de recherche n'ont pas encore été précisément identifiées lors de la collecte initiale, il est vivement recommandé d'informer individuellement les personnes concernées que leurs données peuvent être réutilisées à des fins de

recherche ultérieures, de recueillir leur accord (lorsque la disposition susmentionnée est applicable) et de les inviter à consulter un site web (dont l'adresse est communiquée) qui sera alimenté au fil de l'eau.

La remise d'un document d'information écrit est fréquente. Elle vise à garantir cette nécessaire transparence. De nombreux professionnels de santé ont aussi adopté des portails en ligne pour renforcer et faciliter cette transparence. Ces outils donnent un contrôle accru aux personnes humaines de leurs données. Toutefois, l'efficacité de ces documents est parfois remise en question. Leur complexité et le jargon utilisé peuvent décourager la lecture. Ces documents ne sont pas toujours adaptés à ceux qui doivent donner leur consentement ou être simplement informé. De plus, leur hétérogénéité dans la rédaction complique la tâche des comités de protection des personnes (CPP) pour les études régies par les dispositions du CSP et des règlements européens ou du CESREES pour les recherches n'impliquant pas la personne humaine.

A noter toutefois quelques heureuses initiatives comme le modèle de document d'information mis à disposition par le ministère de la santé et élaboré par l'ensemble des acteurs (associations de patients, établissements de santé, industriels, CPP, DGS, CNIL) en 2021 et la mise en place des collèges de relecteurs de l'Inserm visant à développer un outil qui génère des documents d'information complets (outil Noticeinfobox). Il est primordial de s'inspirer de ces initiatives pour favoriser l'autonomie de consentement des personnes. Créer une notice d'information standardisée et simplifiée sous forme de pictogramme et d'illustration permettrait de réduire les ambiguïtés, de simplifier le langage et garantir que tous les patients reçoivent une information de qualité et facilement compréhensible.

De plus une notice harmonisée permettrait également d'optimiser le travail des comités éthiques et de garantir une meilleure cohérence dans le niveau d'information transmise.

Enfin, en cas de réutilisation de données déjà collectées, dans l'hypothèse où une nouvelle information individuelle ne pourrait être délivrée et que la réutilisation des données n'aurait pas été anticipée au moyen d'un portail de transparence, trois exceptions à la fourniture d'une information individuelle sont prévues par l'article 14.5.b du RGPD pour les traitements mis en œuvre à des fins de recherche scientifique :

1. L'information individuelle se révèle impossible ;
2. Ou exigerait des efforts disproportionnés de la part du responsable de traitement au regard de l'ancienneté des données ou du nombre de personnes concernées ;
3. Ou est susceptible de compromettre ou de rendre impossible la réalisation des objectifs du traitement.

Des mesures appropriées doivent alors être mises en œuvre par le responsable de traitement afin de protéger les droits et libertés des personnes concernées. Dans ce cas, le RGPD, éclairé par les lignes directrices adoptées par le Comité européen de la protection des données (CEPD), prévoit que des mesures appropriées doivent être

prises pour protéger les droits et libertés des personnes concernées, notamment par la mise en place systématique d'une information collective via des canaux de communication appropriés au regard du contexte de l'étude (a minima, sur le site web du responsable de traitement). Une telle dérogation à l'information individuelle est notamment par exemple prévue dans les référentiels applicables aux traitements de données du SNDS (MR-005 à MR-008 et référentiel « ESND »).

L'ensemble de ces règles, contraintes et modalités d'application méritent d'être mieux connues.

Transparence et sécurité des données : Le manque de transparence et les risques non maîtrisés sur la sécurité des données peuvent entraver les projets. Les données de soins, et leur réutilisation dans le cadre de recherche, demandent une vigilance accrue, car touchant à la confidentialité et à l'intimité de chacun. Il est ainsi crucial que ces éléments soient hautement considérés afin de prévenir tout doute auprès des patients et éviter toute inquiétude qui ralentirait la progression de la recherche clinique. Aujourd'hui, les attaques malveillantes et les défaillances sécuritaires, qui touchent principalement les données de soins et peu (pas) les données de soins centralisées pour les recherches, pourraient conduire à interrompre une étude de recherche et surtout fragiliser la confiance nécessaire entre les acteurs. Il est donc crucial de mettre en œuvre toutes les procédures de sécurisation des données, idéalement dans un espace souverain, pour continuer de renforcer la confiance et garantir le succès des recherches.

Recommandations du Groupe “Données” (GT AIS/F-CRIN)

Six axes de recommandations pour l'utilisation des données de santé en recherche clinique :

1. Référencer et documenter les bases de données existantes
2. Garantir la qualité, l'interopérabilité et la standardisation des données
3. Prévoir la reconnaissance et le partage de la valeur dans la collecte et l'utilisation des données
4. Assurer la transparence et la confiance dans la collecte et l'utilisation des données
5. Inciter à la collaboration et au partage de données
6. Assurer la dissémination des modes d'utilisation des données de santé en vie réelle en recherche clinique

Les 18 recommandations ici présentées n'ont d'intérêt que si elles sont suivies d'effet et effectivement appliquées. Leur suivi et leur impact est prévu dans le cadre de la réflexion globale engagée par la démarche initiée par l'Agence de l'innovation en santé et F-CRIN, Infrastructure nationale de recherche clinique, sur « l'évolution des méthodologies d'essais cliniques : nouveaux outils, nouveaux usages et conditions de recours ».

Est également mentionné, en propos liminaires qu'une réflexion est menée sur l'adaptation du cadre juridique aux évolutions des nouvelles méthodologies de recherche clinique.

Axe 1 – Référencer et documenter les bases de données existantes pouvant être exploitées dans l'évaluation de l'innovation en santé

- **RECO 1 : Identifier et référencer au niveau national les bases et entrepôts de données accessibles** avec une caractérisation orientée « utilisateurs/chercheurs ». Ces paramètres de caractérisation devront permettre d'apprécier le périmètre et les usages compatibles, la maturité des données, leur niveau de qualification, le degré d'interopérabilité et les principes d'accessibilité (gouvernance, valorisation, modalités) …
- **RECO 2** : Impliquer les producteurs et utilisateurs de bases et d'entrepôts de données dans la définition du cadre des « **critères minimums** » compatibles à une reconnaissance et prise en compte des projets menés, et dans la définition du « niveau de preuve scientifique et grades des recommandations selon la typologie des données » issus des principes de « l'Evidence-Based Medicine ». En vue de généraliser l'utilisation des données de vie réelle, analyser l'apport des projets menés sur ces bases et entrepôts de données de santé de qualité.

- **RECO 3 : Standardiser la gouvernance et les aspects règlementaires pour le partage de** données issues de plusieurs bases ou EDS afin de favoriser l'accès aux innovations.
- **RECO 4** : Par la **priorisation thématique**, les financements publics de la recherche clinique, type PHRC, doivent inciter au recours à de nouvelles méthodes mettant en œuvre des données de vie réelle, comme la construction de bras de contrôle externe ou l'émulation d'essais cliniques cibles. Une concertation associant l'AIS, le ministère de la Santé, l'Agence nationale de la recherche devrait permettre de définir un cadre d'application pour assurer cette thématisation.

Axe 2 – Garantir la qualité, l'interopérabilité, la standardisation et la pertinence des données au regard de l'utilisation prévue

- **RECO 5 : Définir des critères d'appréciation de la qualité de la donnée et pouvoir vérifier en amont via une phase de tests préalables son utilisabilité au regard de l'utilisation visée.**

  Cette étape de vérification s'apprécie en fonction de trois critères :

  1. La maturité du cycle de vie de la donnée, dont le cycle d'actualisation
  2. L'appréciation du jeu de données vis-à-vis de ses caractéristiques intrinsèques, et le niveau d'adéquation de sa qualité avec les différentes familles d'analyse pour laquelle il est destiné
  3. L'exhaustivité et l'exactitude des métadonnées qui décrivent le jeu de données

  S'appuyer sur un modèle de qualité des données de référence, reconnu par ses pairs, tels que ceux en développement par l'Agence Européenne du Médicament (DQF for EU medicines regulation), s'appuyant sur les dimensions de cohérence, fraîcheur, pertinence, fiabilité et d'exhaustivité pour décrire la qualité et l'utilité des données, permettant d'assurer une compréhension commune, harmonisée et clairement documentée de la qualité des données. Cette démarche d'évaluation de la qualité doit être menée de façon indépendante et standardisée.

  Elle suppose également la poursuite d'expérimentations et de davantage de « preuve de concept » de mise à disposition de jeux de données anonymisées lors d'une phase de tests pour en vérifier l'utilisabilité. L'accès à ces « bacs à sable » se ferait de manière simplifiée au plan juridique.

- **RECO 6** : Assurer **la prise en compte des données de vie réelle pour l'accès à l'innovation** : Intégrer dans la conception en cours du label européen de qualité et d'utilité des données par la Commission Européenne, la mesure de la qualité de données de vie réelles utilisées dans le cadre d'un essai clinique pour

des données auditables par les agences règlementaires pour les accès cliniques.

- **RECO 7** : Mettre en œuvre une **standardisation syntaxique et sémantique compatible avec les standards internationaux** comme vecteur de qualité, permettant de spécialiser et classer les données pour un cadre d'utilisation. Les référentiels sémantiques ne sont pas assez partagés, acceptés et promus par les utilisateurs (société savantes, collège d'experts,). Les utilisateurs devraient être incités pour une implication plus importante sur ces sujets. Ceci permet à la fois d'offrir un niveau minimal de qualité et de proposer une logique de test pour mesurer les dimensions de qualité qui n'auraient pas déjà été assurées.
- **RECO 8** : Développer la recherche sur les méthodologies de mise en qualité des données intégrant l'identification et la neutralisation des biais et l'évaluation du niveau qualité réel des données des données de vie réelle (EDS, Cohortes). Ces recherches pourraient bénéficier d'un financement sur appel d'offre/appel à projets spécifique pour permettre aux équipes expertes sur ces sujets de mener ces travaux. Il pourrait s'agir par exemple d'essais hybrides dont les bras témoins seraient composés de données du SNDS, d'EDS ou des données de santé artificielles.

Axe 3 – Assurer la reconnaissance et le partage de la valeur dans la collecte et l'utilisation des données

- **RECO 9** : En complément des travaux en cours au sein du Comité stratégique des données de santé**, établir au plus vite des principes d'échange et de tarification pour l'accès aux différentes typologies de bases de données avec les partenaires privés.** Ces modalités sont à construire avec les représentants de l'ensemble des parties impliquées (HDH, EDS, établissements de santé, offices de transfert de technologie, investisseurs, grands groupes) pour aboutir à des grands principes (échelle des coûts, cadre règlementaire simplifié) acceptés par tous. Une lisibilité et une transparence sur les conditions de partage seront un critère fort d'attractivité.
- **RECO 10** : Afin de rémunérer le travail sur les données à leur juste valeur, établir pour les cas d'usage identifiés, des **principes de partage de la propriété intellectuelle entre producteur de données et utilisateur** prenant en compte l'apport intellectuel du producteur. Ces règles de partage sont nécessaires lorsque la contribution du producteur des données dans l'acquisition, la qualification, la mise en forme, la documentation et la mise à disposition est effective.

Axe 4 – Améliorer les modes et circuits d'information du patient

- **RECO 11 :** Tout en s'assurant de l'implication des patients et de leurs représentants dans la conception des notices d'information jusqu'à la

communication des résultats des études, **mise en place de notices d'information simplifiées harmonisées** mentionnant systématiquement la possibilité de réutiliser les données et l'adresse web du portail de transparence alimenté au fil de l'eau afin de favoriser les usages secondaires de ces données.

- **RECO 12 :** L'information individuelle des patients lors de la constitution et de l'exploitation des données d'un EDS ou concernant la réutilisation de leurs données (quand un portail de transparence n'a pas été mis en place) pourrait s'appuyer, par exemple, sur l'Espace Numérique Santé sécurisé afin d'assurer aux usagers une information accessible et la capacité de s'opposer à leur utilisation. Cette solution facilitera le travail d'information des porteurs de projets sur données d'EDS. **L'enjeu est aujourd'hui de faciliter le partage des données tout en respectant le choix des individus.**

  Pour cela, le portail monespacesante.fr porté par le ministère de la Santé et la Caisse nationale de l'assurance maladie pourrait être utilisé comme support privilégié d'information et de gestion de l'opposition à la réutilisation des données. Ce système centralisé permettrait de renforcer l'autonomie des patients et de garantir une meilleure gestion et utilisation des données.

Axe 5 – Inciter à la collaboration et au partage de données

- **RECO 13 : Définir des modalités de recevabilité précises par les Autorités des données/sources de données** (ex. données répondant aux normes de qualité de niveau règlementaire garantissant ainsi une « auditabilité » de ces données) en amont d'une évaluation, en vue de garantir une meilleure reconnaissance de la valeur des données de vie réelle utilisées dans le cadre de la recherche clinique. Afin de renforcer l'attractivité au plan international de ces sources de données, définir des conditions d'accès accélérées quand celles-ci ont pour objectif d'être utilisées dans le cadre de la recherche clinique. Des retours positifs sur la prise en compte des données issues des bases de données émis par les autorités de santé françaises, européennes et internationales seraient considérés comme une dimension du label de qualité.

- **RECO 14 :** En lien avec le Comité stratégique des données de santé, **mise en place d'une instance dédiée à la construction de la doctrine relative à la constitution des bases de données.** Sa composition et sa gouvernance garantiraient une bonne représentativité de l'ensemble des acteurs (producteurs de données et monde socio-économique). Elle serait chargée de :

  - Définir une stratégie opérationnelle de données de santé en France basée sur l'usage et émettre des recommandations sur la constitution future des bases/EDS en adéquation avec les enjeux actuels et futurs de la recherche publique et privée (ex. France innovation santé 2030).

  - Assurer plus de cohérence entre les différentes sources (notamment accès précoce, registre, cohorte, génomiques…) et les usages de données de santé.

- **RECO 15 : Proposer un accompagnement personnalisé associant les régulateurs aux porteurs de projets souhaitant mettre en œuvre des essais innovants.** Sont directement concernés la CNRIPH, CNIL, HDH, ANSM. Le groupe de travail (GT) copiloté par l'AIS et F-CRIN va contribuer, en tant qu'amorçage et pré figuration, à définir les besoins et les modalités de réponse à y apporter.

  Ce GT serait chargé de publier des recommandations pratiques au fil des cas d'usage, à l'instar de ce qui est prévu dans le cadre de la phase pilote sur les « essais décentralisés et dématérialisés ». Par ailleurs, ces réflexions pourraient utilement nourrir les travaux de la CNIL dans la cadre de la mise à jour de ses méthodologies de référence.

- **RECO 16 : Définir un modèle économique** (mobilisant des soutiens publics et privés) assurant le financement de la création de bases de données et la collecte de données exhaustives et de qualité tout en définissant de nouveaux indicateurs de mesure de cette qualité.

  Ce modèle prendra en compte la reconnaissance de l'ensemble des acteurs ayant participé au travail de collecte et de validation des données dans la publication/communication des résultats.

Axe 6 – Assurer la dissémination des modes d'utilisation des données de santé en vie réelle en recherche clinique

- **RECO 17 :** Informer et former les membres des organismes amenés à valider le recours à des données de vie réelle dans le cadre des protocoles de recherche clinique. Le cadre de la formation existe aujourd'hui avec le partenariat et la plateforme mis en place par la CNRIPH, la DGS et l'EHESP.

- **RECO 18** : Au sein des formations diplômantes en lien avec la recherche clinique et la recherche sur les données (Master en recherche clinique ; Master en biométrie ; Master en Datascience ; …) intégrer des modules spécifiques sur les méthodologies innovantes d'utilisations des données de santé et leurs limites, et sur les données de santé numérique.

L'importance du "modus operandi" – Mise en action des recommandations du Groupe "Données" (GT AIS/F-CRIN)

Les recommandations formulées dans ce rapport pour assurer « l'utilisabilité » des données de vie réelle dans les activités et les protocoles de recherche, n'ont d'intérêt que si elles sont suivies d'effets. C'est pourquoi le groupe de travail s'est attaché à définir la méthodologie d'application des recommandations, de leur mise en pratique (un mode d'emploi fondé sur le « Qui fait quoi ? ») en partant de quatre principes de base :

1. Assurer la distinction « Données de santé pour le soin/Données de santé pour la recherche » ; Reconnaitre la/ spécificité de l'utilisation des méga-données ou données massives lors de l'évaluation et du suivi en cours d'utilisation des produits de santé ;
2. S'intégrer dans la dynamique globale d'exploitation des données de santé et ses (nombreuses) actions structurantes ;
3. Contacter et informer les organismes/institutions identifiés pour assurer le portage des recommandations et mettre en place si besoin des groupes de travail dédiés.
4. Assurer un suivi collectif à intervalle annuel de l'application, de l'impact, de la pertinence de chacune des recommandations afin de les adapter ou les compléter si besoin.

D'ores et déjà des acteurs et porteurs privilégiés ont été identifiés.

Il est à noter par exemple qu'au plan national la Plateforme de données de santé (Health Data Hub) a engagé un travail de recensement pour les bases qu'il est amené à gérer. Il pourrait être étendu en prenant en compte les initiatives connexes (France Cohortes) afin de disposer d'un catalogue plus complet (Reco 1). Le comité stratégique des données de santé, et la CNRIPH constituent également des effecteurs potentiels, tout comme la CNIL qui va lancer une concertation sur ses référentiels et publier en 2024 une cartographie des EDS dont elle a connaissance.

Ce rapport est la 1ère « brique » d'un travail qui s'inscrit dans la durée. Les priorités, l'organisation seront partagées et définies avec l'ensemble des acteurs du domaine.

↘ MERCI

Pour plus d'informations

camille.schurtz@pm.gouv.fr vincent.diebolt@inserm.fr