HEALTHCARE
DATA INSTITUTE

GÉNÉRATION ET EXPLOITATION DES DONNÉES HOSPITALIÈRES

À VISÉE DE RECHERCHE EN SANTÉ PUBLIQUE ET EN ÉPIDÉMIOLOGIE

MEMBRES DU GROUPE DE TRAVAIL AYANT PARTICIPÉ À L'ÉCRITURE DU POSITION PAPER

Pilotes du groupe de travail

Rémy CHOQUET	Roche	Directeur du Centre des données médicales
Mathieu ROBAIN	Unicancer	Directeur Scientifique Data

Camille BACHOT	Roche	Medical Data Platform Specialist
Manon BELHASSEN	PELyon	Présidente
Isabelle BONGIOVANNI-DELORAZIÈRE	IQVIA	Manager, Real-World Solution
Coralie COURTINARD	Unicancer	Chargée de mission RWD
Timothé CYNOBER	Novartis	Digital Innovation Manager
Romain FINAS	AliraHealth	VP Real-World Evidence
François MARGRAFF	Roche	IT Business Partner Strategic Theme
Sylvaine MAURY	Roche	Directrice Customer Experience
Frédérique MAUREL	IQVIA	Principal, Real-World Solutions
Franck NOGUERO	Pierre Fabre	Directeur Relations Institutionnelles – Stratégie Transverse et Parcours de Soins
Gilles PAUBERT	Cegedim	SVP, Global Head Cegedim Health Data
Samantha PASDELOUP	ELSAN	Directrice partenariats et Développement
Vincent PLANAT	Dedalus	Principal architect – Strategy & innovation
Véronique ROBERT	Unicancer	Data Project Manager
Cheikh TAMBEROU	Cegedim	Directeur de la division RWE Gers Data

SOMMAIRE

RÉSUMÉ

Les systèmes de santé produisent de plus en plus de données, et les exigences en matière d'usage des données de santé augmentent proportionnellement aux besoins de santé publique toujours plus nombreux (COVID, accès à l'innovation, médecine personnalisée). Bien que les sujets des données, de l'IA et du digital soient au centre de beaucoup de communications diverses, parfois érigées au niveau de priorité nationale, la réutilisation des données produites dans le cadre du soin pour la santé publique ou l'innovation reste très complexe. Certains citeront des problématiques techniques, d'autres culturelles, de souveraineté, de confiance ou encore de moyens. Le HDI a décidé de s'emparer du sujet afin de faire un état des lieux de la situation, mieux comprendre la pratique réelle de réutilisation des données de soins, la manière dont les acteurs du soin sont organisés. Nous nous sommes rapprochés de structures de soin afin d'identifier les problématiques rencontrées par ces acteurs qui ont d'abord la mission de soigner. Nous avons construit un outil d'autoévaluation pour ces acteurs afin de les aider à mieux se positionner dans le paysage de la réutilisation des données de santé. Enfin, cette évaluation nous a permis de proposer 32 recommandations dont 20 à destination des établissements de soins et 12 à destination des autorités/décideurs. Ces propositions sont originales, alignées et complémentaires avec d'autres travaux récents (Rapport HAS entrepôts de données de santé, octobre 2022) sur le sujet des entrepôts de données de santé.

Parmi elles, figurent des recommandations très pratiques :

À destination des établissements	Réduire les délais de contractualisation : Bien scinder les discussions relatives à l'investissement d'équipement versus de valorisation des données (mieux connaître ses données, clarifier l'apport externe pour mise à niveau de la donnée, approche règlementaire rapide, modèles par type d'acteur/ type de projet, gestion de la PI, sous-traitance, faisabilité plus rapide)
À destination des autorités de santé et décideurs	Adapter les modèles de financement public pour les producteurs de données utilisables dans le cadre de l'évaluation des produits de santé et de santé publique. Par exemple, une reconnaissance des indicateurs RNIPH pour tous les producteurs de données participant aux études. Un mécanisme de type ROSP pourrait être aussi adapté pour les producteurs de ville.
À destination des autorités de santé et décideurs	Labelliser des circuits de partage et d'exploitation de données de confiance (public ou privé), thématiques (santé publique, par maladie etc.), certifiable, auditable, et couvrant les dimensions technico-réglementaires et leur usage scientifique

INTRODUCTION

LES DONNÉES DE SANTÉ SONT NÉCESSAIRES À LA SANTÉ PUBLIQUE

Les données de santé sont des données collectées au cours de diverses activités médicales ou non, qui peuvent être relatives aux déterminants généraux de santé, et à la santé d'une personne, d'un groupe de personnes (couple, famille, quartier, ville, région, ethnie, pays, etc.) ou de populations (santé publique, santé au travail). Ces données peuvent être utilisées pour le suivi et l'évaluation des systèmes et politiques de santé, des produits de santé, mais aussi dans le domaine de la prévention ou bien pour créer des dispositifs médicaux de diagnostic de nouvelle génération. Elles sont nécessaires au pilotage du système de santé, notamment des crises sanitaires. La majeure partie des données de santé utilisées en Santé publique sont des données collectées spécifiquement pour répondre à des questions de recherche. Leur coût de collecte est très important, parce qu'une saisie des données dans un cadre structuré est nécessaire, et pose bien souvent la question de la soutenabilité de ces approches sur le long cours.

Les données issues de l'activité de soin des établissements de santé sont aujourd'hui perçues comme une richesse sous-exploitée. Face à l'augmentation du volume de données, leur diversité, leur exploitation et leur valorisation sont une réelle opportunité afin de moderniser les approches classiques de santé publique et d'améliorer la soutenabilité de celles-ci. Cependant, il demeure de nombreux challenges même si nous pouvons tous observer que la dynamique est lancée, voire s'accélère.

La base du Système National des Données de Santé SNDS, gérée par la CNAMTS, est historiquement la base de données nationale de remboursement des actes et prescriptions en médecine de ville et des éléments d'activités des établissements de santé. La base du SNDS, décrite comme l'une des plus grandes bases de données de santé exhaustive, n'aborde pas d'éléments de morbidité et ne comporte pas les données relatives aux résultats cliniques des actes et examens répertoriés. L'accès pour la recherche et l'évaluation à ces données peu médicalisées poursuit sa mutation pour faciliter le rapprochement indispensable à des bases enrichies de données médicalisées. Cela permettra de mieux répondre aux nombreuses questions de recherche en épidémiologie et en santé publique, et de favoriser les études, recherches ou évaluations sur les traitements dispensés lors des parcours de soins de l'ensemble des bénéficiaires du système de santé, les innovations, la surveillance de certaines pathologies, la connaissance des dépenses de santé.

Les grandes cohortes françaises (France cohortes, Inserm, Unicancer) déployées dans plusieurs groupes de pathologies impliquent un travail collaboratif pour permettre de centraliser et structurer des données longitudinales et générer des connaissances nouvelles. Les structures de recherche publique, les regroupements d'établissements de santé, mais aussi des mutuelles et parfois des acteurs industriels permettent ensemble de mettre en place et de suivre de façon pérenne ces données de cohortes. Cela n'est possible qu'avec une gouvernance partagée et cohérente avec l'ensemble des acteurs impliqués. Leur apport est majeur dans la description et l'étude des facteurs liés à la survenue d'événements de santé, mais reste par construction largement contraint par le « temps long » des cohortes et par les contraintes définies lors de leur conception.

Les registres de morbidité populationnels généraux ou spécifiques (DGOS, INCa, Agence de la Biomédecine) sont définis depuis 1995 comme un recueil continu et exhaustif de données nominatives sur des événements de santé dans une population géographique définie. Les 67 registres identifiés, en plus des 73 registres spécifiques sur les maladies rares, apportent un niveau d'indicateurs de morbidité strictement descriptifs et ont une valeur de référence sur les populations concernées. Une réflexion stratégique est menée depuis plusieurs années pour mieux encore prendre en compte ces données et les intégrer dans une démarche de recherche en épidémiologie plus large que celle uniquement en lien avec ses objectifs premiers. La Direction générale de la Santé, le Haut Conseil de la Santé publique ont émis en 2021 des recommandations pour valoriser la pertinence et l'utilité des registres[1]. Cela passe par l'homogénéisation de la qualité pour favoriser la valorisation des données, notamment en lien avec d'autres sources de données, mais aussi développer d'autres finalités comme la détection des signaux faibles. Les notions relatives aux modalités de prise en charge, tels les actes et prescriptions restent cependant difficiles à rapprocher des événements cliniques suivis.

Les bases de données de santé gérées par les industriels sont le plus souvent restreintes en termes de patients et centrées sur un produit de santé (médicament, dispositif médical ou technologie de santé). Ces bases rarement représentatives, mais avec un niveau qualitatif important ont tendance à s'ouvrir au partage. Leur approche spécifique et très centrée « produit » est une limite pour permettre leur utilisation dans la recherche en épidémiologie et Santé publique.

Les bases de données issues des panels de médecins libéraux et officines de ville (Cegedim, IQVIA) reposent sur des outils de captation des données médicales (logiciels médicaux) ayant bénéficié d'une standardisation par une certification HAS, et des outils de dispensation des produits de santé en officines de ville communicants. Ces démarches permettent une approche partielle pour la recherche du fait que la médecine de ville est majoritairement concernée. Ces bases « propriétaires » des entités qui gèrent/centralisent et exploitent ces données, devront à terme pouvoir communiquer avec des bases de données hospitalières structurées et de qualité pour mieux servir les ambitions de recherche en épidémiologie.

Depuis 2018, la France poursuit son engagement[2] dans une politique et un soutien actif à la valorisation des données dans le domaine de la Santé. Pour générer des connaissances nouvelles

en vue de développer des pratiques de soins optimales, l'exploitation des données de santé ne se conçoit qu'avec la participation de l'ensemble des établissements de santé au bénéfice de tous. Le développement du partage de données maîtrisé avec une notion forte de sécurisation et de protection des citoyens reste des principes intangibles. De nombreuses initiatives nationales et locales sont menées tant dans la constitution et structuration d'entrepôts de données de santé (DGOS, BPI), que dans les recherches de méthodologies d'analyse adaptées et spécifiques dont l'IA. La volonté est clairement d'être à même de traiter et exploiter à leur juste valeur ces données aux formats et finalités hétérogènes, et quantitativement très importantes.

Les projets réalisés donnent lieu à de nombreuses publications d'articles scientifiques et éclairent également les Autorités de Santé dans leur prise de décisions. Ainsi, c'est en France que la plus grande étude sur l'efficacité des vaccins anticovid a été réalisée, avec 22 millions de personnes incluses. Elle a permis de prouver que la vaccination contre le Covid-19 réduisait de 90 % le risque d'hospitalisation et de décès chez les plus de 50 ans. Ces données de santé ont également été utilisées dans le VIH, avec depuis 2017, la réalisation par EPI-PHARE du suivi annuel de l'évolution de l'utilisation de l'association Emtricitabine/Ténofovir/Disoproxyl ou génériques pour une prophylaxie préexposition (PrEP) au VIH.

À l'international, la coexistence de ces différents types de sources de données médicales prend en compte les spécificités nationales avec leur organisation et leurs limites. Malgré les différences d'organisation du système de santé et des caractéristiques des populations nationales, il semble plus naturel de faire communiquer les bases de données publiques entre elles. Cependant, la complexité de considérer les données médicales et thérapeutiques hospitalières trouve les mêmes écueils qu'en France, même si différentes initiatives multipays commencent à voir le jour. Un travail important sur la standardisation et l'interopérabilité reste un prérequis pour répondre à des questions de recherche communes dans le domaine de l'épidémiologie et de la Santé publique. La Commission européenne, via l'opérateur français Health Data Hub (HDH), porte la coordination de l'Espace Européen des Données de Santé (EEDS) indispensable en termes d'infrastructure. Les objectifs planifiés et réalistes restent à mettre en œuvre entre les différents pays concernés tenant compte des spécificités des données hospitalières de chacun.

Au front de la prise en charge, les établissements de santé sont un maillon essentiel dans la chaîne de la recherche en santé publique et en épidémiologie. En effet, la réutilisation des données de santé issues de la prise en charge est de plus en plus importante pour mieux comprendre les parcours de soins, l'efficacité et la sécurité des produits de santé, et ainsi éclairer les pouvoirs publics dans la prise de décision plus rapidement. Les établissements de santé ont cependant un niveau de maturité hétérogène dans la réutilisation de leurs données de soins pour leur propre besoin ou bien avec des établissements ou des groupements hospitaliers déjà conscients des enjeux sur le sujet, et des établissements n'ayant pas les ressources pour s'engager dans cette valorisation des données. Les données médicales hospitalières dépendent par ailleurs de systèmes d'information et d'outils hétérogènes peu communicants entre établissements.

Récemment, un appel à projets (AAP) d'un montant de 50M d'euros a été lancé par le ministère

de la Santé, appuyé par le HDH, afin d'aider les établissements à s'équiper et à embrasser une démarche locale de structuration de leurs données de soin[3]. Bien que cet AAP soit un élément clé pour accompagner une démarche nécessaire, nous pensons qu'il est aujourd'hui assez difficile de comprendre d'où les établissements partent, du niveau de maturité de cette démarche à ce jour, et des leviers qu'il serait nécessaire de déployer afin d'assurer une véritable mise à l'échelle des investissements prévus. Il reste par exemple assez difficile de savoir si le potentiel de réutilisation des données de soins via des entrepôts de données de santé est optimal, si les hôpitaux se sont dotés d'une stratégie sur la réutilisation de leurs données ou si les compétences sont présentes…

Ce position paper vise à faire un état des lieux des enjeux pour un établissement de santé, à faire état de la maturité d'un petit nombre probablement pas représentatif d'établissements de santé sur le sujet, et de proposer des recommandations afin d'accélérer la mutation en cours de notre outil commun de production de données pour la santé publique.

PARTIE 1

DE NOUVELLES DONNÉES ET DE NOUVELLES APPROCHES SE DÉVELOPPENT

LA DONNÉE DE VIE RÉELLE DOIT RACONTER UNE HISTOIRE, CELLE DE LA PRISE EN CHARGE D'UN PATIENT

Historiquement, la recherche observationnelle consistait à planifier la collecte dans un cadre *ad hoc*, souvent réalisée par des centres investigateurs **reconstituant manuellement des bases en «lisant» les informations disponibles là où elles se trouvaient** (dossier médical, compte-rendu de consultations, d'interventions, d'hospitalisation, de biologie, d'imagerie, d'anatomopathologie, relevés administratifs). Les conditions de recueil des données et surtout de leur contrôle qualité étaient ici proches de celles réalisées dans le cadre d'une recherche expérimentale, un essai clinique. Les principales limites de ce type de collecte touchent aux aspects de coûts (humains) et de délais (collecte manuelle).

La réutilisation automatisée issue des logiciels du soin (logiciels de biologie, d'imagerie, dossier, patient, dossiers administratifs…) offre bien sûr **un accès potentiellement plus rapide à un volume de données important**, issues de **plusieurs sources** (biologie, imagerie…), mais également **tout au long du parcours du soin**, de l'hôpital à la consultation en ville, et maintenant au domicile (hospitalisation à domicile, prestataires de services, dispositifs médicaux communicants, objets connectés).

Les données recueillies en condition réelle de prise en charge peuvent contribuer à plusieurs usages :

→ **La recherche**, comme source de données dans le cadre de différents types d'approche :

- **Approche populationnelle** : Études de santé publique (épidémiologie, prévention, surveillance et veille sanitaire, histoire naturelle de la maladie, économie de la santé) (ex. : registres, cohortes, observatoires, autres…)

- **Approche parcours et organisation des soins** : Études de qualité des soins et des parcours de soins (amélioration des soins, évaluation de la performance entre chirurgiens), évaluation de l'efficience et l'efficacité de l'organisation des soins.

- Approche **produit de santé** (médicaments, dispositifs médicaux, thérapies digitales) : études et activités liées au développement des produits de santé, identification et inclusion des patients en recherche clinique, études des bénéfices, des risques, du coût-efficacité ou encore de l'utilité perçue par les patients.

→ Le **développement** de méthodologies et innovations technologiques telles que l'aide à la décision, IA, méthodologie d'essais cliniques, etc. pour prévenir, prédire ou personnaliser les prises en charge.

→ Le **pilotage**, qu'il soit **stratégique** (financier, recrutement médical, développement de nouvelles offres de soins, projet médical de l'établissement, etc.) ou **opérationnel** (gestion des flux d'activité de l'établissement ou par service, occupation des lits, blocs opératoires, efficience de l'établissement, des services, codage, organisation des services, etc.)

L'enjeu est donc de pouvoir mettre à disposition un ensemble de données qui reflète l'histoire médicale réelle de prise en charge du patient et de son environnement (écosystème de santé).

Ainsi, diverses données peuvent ainsi être utilisées dans des études de recherche en santé publique :

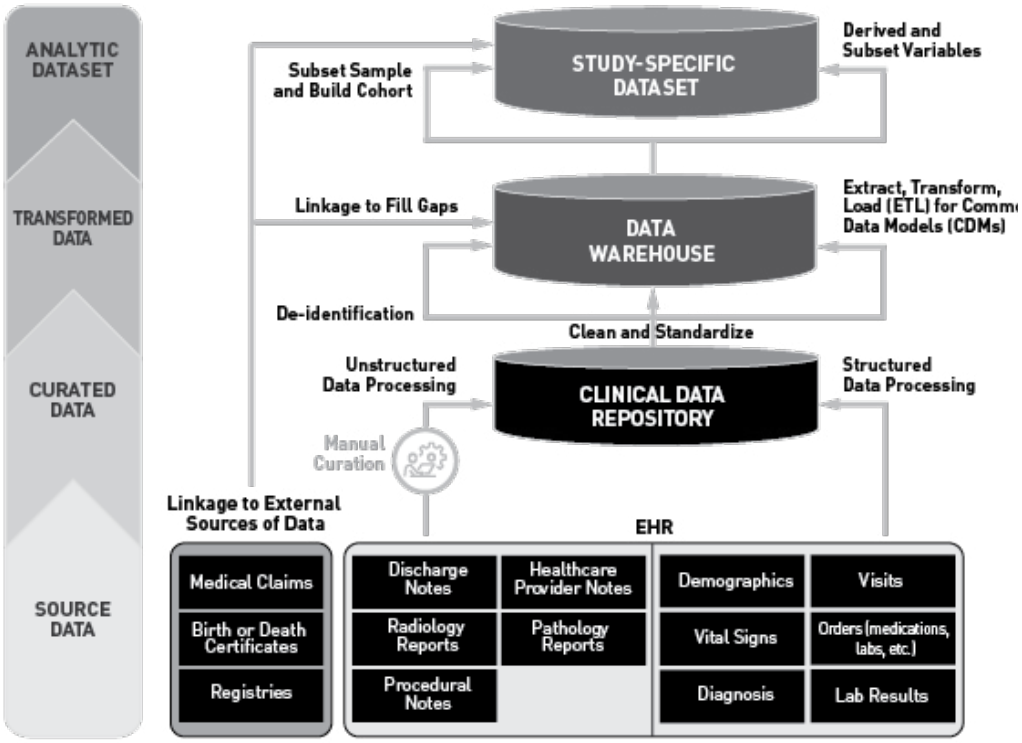
→ **Le profil des patients** étudiés (profil sociodémographique, antécédents, comorbidités, patrimoine génétique…)
→ **L'histoire de leur maladie** (progression des tumeurs, dégradation progressive d'une fonction d'un organe) voire les caractéristiques de certaines maladies (par exemple les mutations génétiques d'une tumeur)
→ **Les actes et conclusions diagnostiques** pratiqués avec les contenus associés (imagerie, biologie, tests fonctionnels, anatomopathologiques, altérations génétiques)
→ **L'histoire des traitements mis en œuvre** (médicaments et dispositifs médicaux, radiothérapie, actes de kinésithérapie, changement de régime alimentaire…) : leur dosage, leur fréquence, l'impact observé sur la maladie
→ **Les effets secondaires éventuels** des traitements et les **médications associées**
→ **La qualité de vie** ou **l'expérience de la prise en charge** du patient
→ **Les ressources mobilisées, humaines ou financières**, les coûts associés payés par les assureurs, mutuelles, et le patient lui-même

Plus on dispose d'informations fiables et pertinentes sur l'histoire des patients, plus les conclusions seront robustes. Il s'agit donc de les trouver et d'en trouver de bonne qualité pour qu'elles soient exploitables. À titre d'exemple, les données cliniques sont essentielles pour contextualiser les échantillons et images issues d'analyse de prélèvements de sang ou de tissus afin de pouvoir développer de nouvelles cibles thérapeutiques.

RÉUTILISER DES DONNÉES DU SOIN POUR LA RECHERCHE, UNE ACTIVITÉ À PROPREMENT PARLER

On ne peut cependant pas résumer la réutilisation de données existantes à la possibilité d'accéder à des bases d'information consolidées. **C'est avant tout une chaîne de transformation** qu'il faut opérer, de la source (le logiciel, le dispositif médical) à la mise à disposition des utilisateurs. L'Agence américaine du médicament, la *U.S. Food and Drug Administration* (FDA), dans ses recommandations à l'industrie en 2022[4] décrit trois étapes de transformation nécessaires pour passer d'une donnée « brute » à une donnée « exploitable » :

Figure 1 : chaîne transformation de la donnée, FDA 2022



1. **La phase d'acquisition** des données issues de logiciels selon leur format (compte-rendu, fichier d'imagerie, données de biologies) pour les consolider dans un » dépôt » archivé autour d'un identifiant commun. Cette phase comprend une première phase de contrôle qualité qui aboutit à une sélection de dossiers en fonction des données disponibles.

2. **Une phase de standardisation** pour rendre la donnée homogène (mise sous un modèle commun de données (OMOP[5,6] ou FHIR[7]) par exemple qui sont maintenant des formats reconnus] et de **nettoyage** (contrôler la plausibilité des données et censurer des patients ou des données aberrantes). Les données sont **ensuite anonymisées et/ou pseudonymisées** et stockées dans un entrepôt de données de santé (EDS).

3. **Une phase d'extraction** qui permet ensuite de constituer des jeux de données spécifiques selon les exigences de protocoles de recherche ou d'évaluation (critères de sélection des patients, temporalité des données, données à obtenir). Pour éviter toute forme de fuite de données individuelles, certaines bases proposent de réaliser les études (analyses statistiques) dans un environnement spécifique et sécurisé offrant des capacités de calcul adaptées. En France, travailler sur les données issues de l'assurance maladie (SNDS) ne peut se faire que dans l'environnement de la CNAMTS par des personnes habilitées et/ou dans des espaces homologués et conformes à un référentiel[8] (système fils).

D'une logique historique de collecte *ad hoc* et souvent manuelle (collection de données primaires), les méthodes de recueil ont donc fortement évolué vers une réutilisation des données existantes produites dans le soin, mais cela suppose la maîtrise complète de la chaîne de transformation.

QUELQUES EXEMPLES D'ENTREPÔTS DE DONNÉES DE SANTÉ HOSPITALIERS IMPORTANTS EN FRANCE

Cela fait plus de 10 ans qu'un projet mixte, hospitalo-universitaire est développé au sein du CHU de Rennes : le projet eHop. Le projet est aujourd'hui présent dans 18 établissements de santé français. 5 millions de patients représentent le grand ouest français dans le cadre de HUGO par exemple[9].

Depuis quelques années, l'Institut Imagine développe un entrepôt de données médicales avec divers objectifs de soutien à la recherche de l'institut et de partenaires externes[10]. S'appuyant sur des technologies avancées de fouille de textes médicaux, l'outil permet entre autres d'identifier des phénotypes comparables entre différentes prises en charge ou encore de mettre en place des cohortes de patients très rapidement.

Dans le domaine de l'Oncologie, Unicancer met en place, depuis plusieurs années, des entrepôts de données de santé se basant sur des données de soin (programme ESME, observatoire ODH, plateforme Weshare). Ces entrepôts de données s'appuient sur les données de soins hospitalières générées par les soignants, mais aussi par les patients. Depuis 2017, l'Assistance publique Hôpitaux de Paris a mis en place un entrepôt de données de santé[11] rendant compte de la prise en charge de plus de 8 millions de patients hospitalisés à l'APHP dans un des 39 établissements de l'APHP. En janvier 2023, c'est plus de 175 recherches qui ont été réalisées sur cet entrepôt[8].

Des acteurs privés proposent également l'accès à un réseau qui centralise des données de santé anonymisées en provenance de différents acteurs de la chaîne de soins (établissements hospitaliers, pharmacies, etc.). Ces réseaux s'inscrivent dans des initiatives de recherche et développement en permettant d'établir un lien avec les établissements sources pour engager des essais cliniques.

FAIRE COÏNCIDER LES DONNÉES PRODUITES DANS LE PASSÉ AVEC LES BESOINS ACTUELS ET À VENIR DE LA RECHERCHE

Historiquement, les jeux de données *ad hoc* (données primaires) étaient collectés pour répondre point par point aux besoins spécifiques décrits dans un protocole de recherche, la collecte pouvant être **rétrospective,** mais également **prospective**, c'est-à-dire que les données manquantes ou nécessaires étaient systématiquement collectées dans l'exercice de soins futurs.

Pour répondre aux besoins, il faudra donc penser les entrepôts de données de santé (EDS) pour pouvoir continuer à répondre à cette double approche : Faire avec ce que l'on a d'un côté, mais également pouvoir compléter cette information dans le futur. Cela implique non seulement **une capacité d'anticipation** (par exemple pour organiser la digitalisation des lames de biopsies, l'harmonisation des données d'imagerie par exemple) et **une agilité dans la collecte**, offrant non seulement la possibilité d'interroger de manière rétrospective les données (faire avec ce que l'on a), mais également pouvoir compléter et enrichir les données de manière prospective (i.e. sur les patients à venir). Pour cela, il est nécessaire d'intégrer très tôt les besoins des équipes réalisant des recherches sur données de santé dans les processus de décisions IT et digitaux de l'établissement de santé. À titre d'exemple, cela offrirait la possibilité :

- D'interroger les patients pour contribuer à l'évaluation des résultats en santé ou de leur expérience (PROMs et PREMs de plus en plus souhaités par la HAS) ;
- De suivre les patients à distance (données issues de l'activité de télésurveillance) ;
- De pouvoir faire des actes diagnostiques *a posteriori* sur des populations passées comme par exemple la réalisation des tests génétiques sur des prélèvements de tumeurs ou de sang conservés dans des biobanques.

Mettre en place des designs d'études ambispectives, c'est-à-dire disposer de toutes les données du passé tout en recueillant de manière prospective des données complémentaires et multimodales, **devient ici un enjeu majeur d'attractivité dans la compétition mondiale** des bases de données de santé, avec à la clé :

- La souveraineté d'évaluation des innovations ;
- L'innovation sur la donnée comme des IA prédictives et préventives et bien sûr la médecine personnalisée.

FAIRE ÉMERGER DES LEADERS EUROPÉENS DE LA GÉNÉRATION DE DONNÉES

La concurrence est d'ores et déjà importante entre des acteurs internationaux notamment américains (Cleveland Clinic, John Hopkins, Mount Sinaï), israéliens (Maccabi) et chinois, mais également acteurs spécialisés (TrinetX, Flatiron, Concert AI…). Une analyse systématique de PubMed sur une période de14 mois (de janvier 2021 à fin février 2022), montre que concernant le cancer du poumon, *i.e.* le plus meurtrier, avec près de 1300 essais cliniques de phase II et III en cours, 120 publications sur 60 bases existantes ont été menées. Une analyse plus fine de ces bases montre que seulement deux proposent une offre multimodale combinant plusieurs sources comme imagerie, dossier médical, séquençage du génome, PRO et données économiques. L'émergence d'acteurs paneuropéens de la donnée est aujourd'hui centrale pour répondre aux enjeux de souveraineté des prises de décision, d'attractivité de la recherche et *in fine* d'accès aux innovations. Les aspects réglementaires (EEDS, RGPD), technologiques (interopérabilité) s'étant clarifiés depuis, la question est donc maintenant scientifique et industrielle.

LA SÉCURISATION DES ENTREPÔTS : UN ENJEU À NE PAS SOUS-ESTIMER

733 incidents sur des systèmes d'information pour la santé ont été déclarés en 2021[12] en France (augmentation de 100 % par rapport à 2020) dont 65 % pour des établissements de santé. Pour la moitié d'entre eux, tout ou partie des données des applications de la structure n'étaient plus accessibles. 11 % de ces événements ont mis en danger le patient.

Devant ce problème, la CNIL a récemment publié un référentiel sur les entrepôts de données[13] qui liste au chapitre 10 l'ensemble des exigences de sécurité techniques et organisationnelles à mettre en place et valider pour tout déploiement d'EDS.

Ce référentiel couvre un espace technologique pour lequel de multiples solutions existent tant au niveau de l'infrastructure, des systèmes d'exploitation que des applications unifiées de gestion de sécurité (identification, supervision, log et traces, etc.). L'enjeu est alors de faire les choix techniques appropriés et de les intégrer pour ne laisser aucun maillon manquant dans cette chaîne complexe, ce qui exige des expertises dédiées avancées et au fait des dernières évolutions dans le domaine de la cybersécurité.

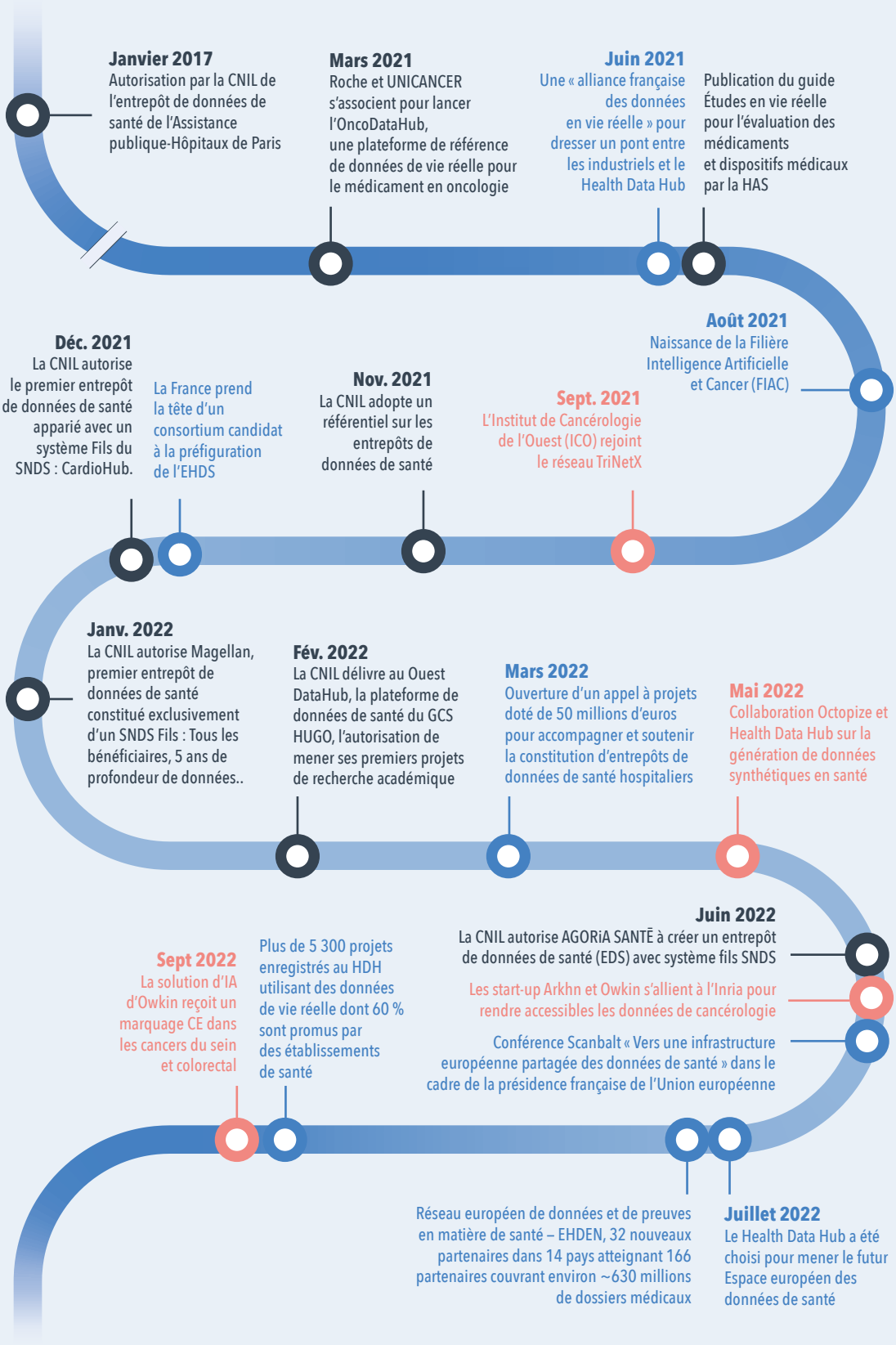
DES MODALITÉS D'EXPLOITATION D'ENTREPÔTS QUI SE DIVERSIFIENT VERS LE MULTISITE

Au-delà des entrepôts locaux à l'établissement de santé, les initiatives multisites permettent de démultiplier le volume de données disponible devant un accélérateur pour la recherche, l'analyse populationnelle incluant les données de vie réelle et le développement de l'intelligence artificielle. Trois types distincts d'entrepôts multisites apparaissent (à des degrés divers de maturité).

- Les multisites à centralisation régionale ou nationale tels que mentionnés précédemment.

- Les multisites de données fédérés comme les projets européens EHDEN ou DARWIN qui eux privilégient des agrégations contributives de statistiques sur plusieurs sites. Dans ce cas, chaque établissement communique un catalogue de ses datasets (métadonnées ne contenant aucune donnée à caractère personnel). Les chercheurs peuvent les consulter et opérer des agrégations de statistiques.

- Les multisites d'apprentissage fédérés[14] quant à eux privilégient la division de la tâche d'entraînement d'un algorithme sur plusieurs machines distantes pour ensuite procéder à une reconsolidation des paramètres locaux sur un serveur central. Dans ce cas, c'est l'algorithme qui se « déplace » vers la donnée et non l'inverse comme dans le modèle classique d'entraînement centralisé.



DATES CLÉS AUTOUR DE LA RÉUTILISATION DES DONNÉES DE SOINS

Réglementaire · Écosystème · Cas d'usage/ Techno

L'exemple canadien de la « Unity Health », Saint Micheal's Toronto Hospital

Le docteur Muhammad Mamdani, de l'hôpital Saint Michael de Toronto, a mis en place une unité de recherche au sein de son établissement pour exploiter les données produites dans le cadre du soin[15]. L'équipe, composée d'une cinquantaine de membres, travaille principalement sur des projets initiés en interne. L'entrepôt est nourri de 2 000 types de données en provenance de 9 systèmes d'information, et recense plus de 300 000 patients uniques.

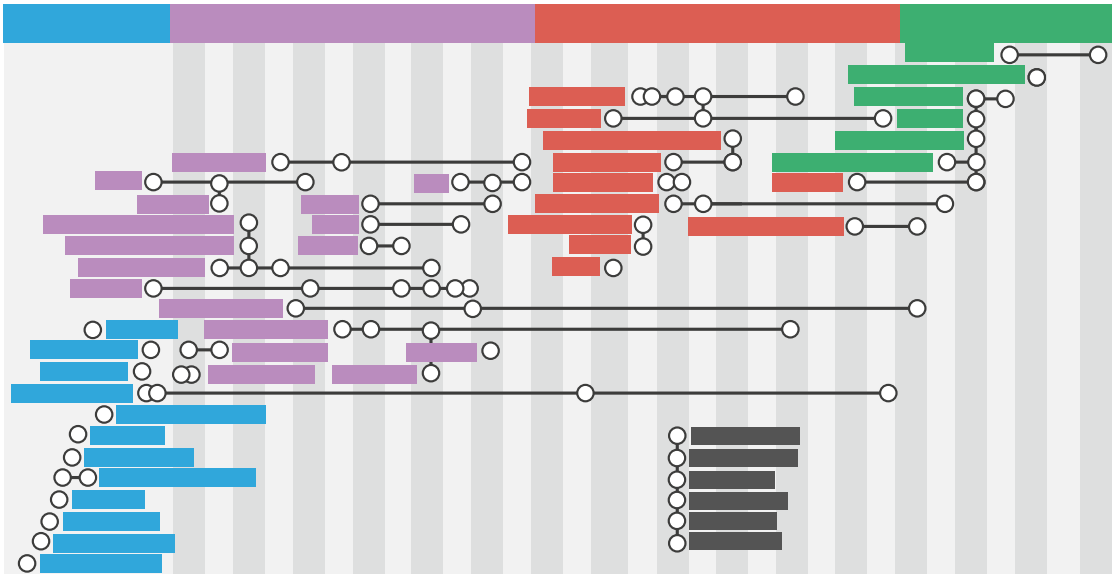


Figure 2 : exemple du pipeline de projet de l'équipe du département Data Science and Advanced Analytics

En moyenne, l'équipe gère 10 à 15 projets en parallèle dont les durées vont de 3 à 6 mois en moyenne. Plusieurs travaux de l'équipe ont été publiés, dont une méthode de validation d'un modèle de prédiction des septicémies[16], ou encore l'évaluation des solutions fondées sur l'apprentissage machine en santé[17].

PARTIE 2

RÉUTILISATION DES DONNÉES HOSPITALIÈRES POUR LA RECHERCHE ET L'INNOVATION, UN PREMIER ÉTAT DES LIEUX

La France a montré ces dernières années sa volonté d'organiser à grande échelle la réutilisation de ces données de santé et de mettre à disposition des acteurs de l'écosystème des outils performants de réutilisation et de partage (experts, outils de calculs et d'analyses, infrastructure technique sécurisée) via la création notamment de la Plateforme de donnée de santé, ou Health data Hub. En amont de tout partage, chaque établissement de santé doit pouvoir être en mesure d'évaluer sa propre «capacité» de réutilisation ou «readiness» de ses données pour des projets de recherche internes ou des projets interétablissements. Un questionnaire standardisé a donc été construit afin de mieux appréhender la capacité d'un établissement à partager des données de soin pour la recherche.

À la fin de l'année 2021, 2984 établissements de santé comprenant 1342 hôpitaux publics, 661 établissements privés à but non lucratif et 981 cliniques privées sont référencés en France[18]. Chaque établissement dispose d'une organisation des systèmes d'information orientée sur la prise en charge des patients au sein des services ainsi que le pilotage opérationnel et fonctionnel. Le potentiel de réutilisation dépend des missions de l'établissement. Dans le cas des CHU ou des CLCCs, la recherche étant une des missions de ces types d'établissements, la réutilisation des données de soin et leur valorisation apparaissent comme un levier de croissance scientifique majeur. Qu'en est-il des autres établissements, qui disposent de données médicales similaires[I] ?

Le groupe de travail a proposé un questionnaire avec les axes suivants :

- typologie de l'établissement,
- stratégie vis-à-vis de la réutilisation des données,
- pratiques vis-à-vis de la réutilisation (usage et partage des données) ainsi que les ambitions,
- moyens technologiques et humains soutenus par des sources de financement,
- modalités actuelles de gouvernance encadrant les données de soin disponibles au sein de l'établissement.

I. Les données médicales couvrent aussi bien les données documentées dans le dossier médical informatisé du patient telles que les consultations, les comptes rendus d'imagerie ou de biologie, les comptes rendus opératoires ou d'hospitalisation ainsi que les données structurées du programme de médicalisation des systèmes d'information.

Un outil a été développé sous la forme d'un questionnaire avec l'objectif de pouvoir proposer un score multidimensionnel. Ce score comprend une quarantaine d'items, pondérés en fonction des réponses sélectionnées.

Au cours de l'été 2022, une dizaine d'établissements volontaires ont été contactés pour tester ce questionnaire de façon anonyme et confidentielle. Pour un même établissement, l'administration de ce questionnaire a nécessité de rassembler autour de la table différents acteurs de l'hôpital comme les cliniciens, les médecins d'information médicale, la direction de la recherche clinique, la direction des systèmes d'information ou le cas échéant, la direction des données. Parfois, l'administration de ce questionnaire a servi à établir un langage commun entre les différents acteurs de l'établissement ou à renforcer une vision commune autour de la réutilisation de la donnée de soin.

RÉSULTATS

La grande majorité des établissements répondants (70 %) était des établissements publics ou participants au service public hospitalier comme les établissements de santé privés d'intérêt collectif (ESPIC). Tous avaient une activité MCO (médecine, chirurgie, obstétrique et odontologie) et seuls trois établissements publics comportaient une activité mixte de type de soins de suite et réadaptation (SSR) et/ou de psychiatrie (PSY).

50 % des établissements de santé ont formalisé une stratégie au travers de l'existence ou la création d'une feuille de route dédiée ou d'une politique de gouvernance relative à cette réutilisation. Favoriser l'accessibilité, assurer la sécurité, maîtriser la qualité a été retenue comme **des objectifs de réutilisation.**

Les domaines de réutilisation de ces données générées par les offreurs de soins sont catégorisés de la manière suivante :

- études de santé publique,
- étude de qualité des soins,
- études liées à l'évaluation et/ou au développement de produits de santé,
- activités relatives au développement de nouvelles méthodologies et d'innovations technologiques,
- pilotage stratégique et opérationnel de l'établissement

Les axes de progression souhaités couvrent **les études pour l'évaluation des produits de santé**, la **qualité des soins** et le développement d'**innovations technologiques**.

Pour la très grande majorité des établissements **(80 %), l'expérience de l'usage des données** se fait au travers d'une politique ou démarche de partage dans un contexte d'utilisation encadré par un tiers (utilisateur unique).

La quasi-totalité des répondants encadre la réutilisation sous forme de contrat de partenariats dont **50 % avec des acteurs privés.** Les autres types de partenariat restent dans le cadre académique (universitaire ou public). Bien que la démarche de réutilisation soit effective, aucun des répondants n'a reconnu qu'une offre d'analyses de données par des tiers était promue de façon proactive par l'établissement.

Concernant les moyens actuels disponibles au sein de l'établissement, **50 % des répondants ont déclaré disposer d'un pôle dédié et d'une infrastructure type Entrepôt de données de santé (EDS)** comme moyens humains et techniques dédiés à l'exploitation des données de soin (structuration, utilisation, accès).

En complément, la mise en place d'un comité spécifique de revue des projets de réutilisation est effective chez **seulement 33 % des établissements**, ce qui évoque la structuration d'un modèle de gouvernance.

Dans la gestion opérationnelle de l'établissement comme, par exemple, pour optimiser les flux de patients, l'utilisation d'outils **d'aide à la décision** alimentés par des données de soins reste marginale (**30 % des répondants**).

L'essentiel des données de soins se trouvant dans les **comptes-rendus textuels** rédigés par les cliniciens, il apparaît que **plus de la moitié** des établissements interrogés ont mis en place des traitements avancés de ces informations écrites à l'aide de techniques liées à l'**intelligence artificielle** pour valoriser la réutilisation de ces données. La validité des données ne fait pas l'objet de mesures ou indicateurs standardisés (80 %) pour soutenir la maîtrise de ces données par les utilisateurs.

Pour soutenir ces démarches, il est intéressant de noter que l'**expertise spécifique** relative à la réutilisation des données est présente dans **6 établissements sur 10,** mais avec une **taille d'équipe assez réduite** entre 5 et 15 personnes généralement. Autour de la réutilisation de données, par les métiers impliqués, le délégué à la protection des données reste le métier majoritaire mentionné (60 %) juste avant le chef de projet (40 %) puis arrivent les métiers plus techniques (comme data scientist ; data ingénieur, data manager et data architect).

Pour soutenir la réutilisation des données, la majorité des établissements s'appuient sur des financements de sources diverses (3 en moyenne) incluant des financements internes, des partenariats de recherche (financement public) et des contrats de collaboration (financement privé). Néanmoins, ces **sources de financement ne semblent pas assez pérennes** pour une grande majorité des répondants (70 %).

Malgré un engagement autour de la réutilisation, **la gouvernance reste un chantier à formaliser** puisqu'il apparaît que peu d'établissements fournissent une formation systématique liée à l'usage de ces données (10 %). Le volet lié à la transparence auprès des patients est amorcé puisque 40 % des établissements déclarent disposer d'un portail dédié afin que les patients puissent faire valoir leurs droits vis-à-vis de ces réutilisations. Enfin, une majorité des répondants (60 %) a confirmé avoir été formée au respect des règles sur la protection des données.

Ce premier travail a permis de mettre en évidence certains facteurs non discriminants sur les dimensions relatives à la typologie de l'établissement et à l'usage de données. Ce questionnaire, s'il est utilisé plus largement et dans le temps, pourra être un outil puissant de pilotage d'évolution d'usage des données produites dans le cadre de processus de soins pour la recherche.

Le centre hospitalier de Valencienne (CHV) et le Pôle Santé République de Clermont-Ferrand ont accepté d'apporter leur témoignage, figurant en annexe 2, en marge du remplissage du questionnaire.

Parmi les initiatives lancées par ces deux établissements, le CHV a mis en place en 2022 un comité d'évaluation des projets afin de les prioriser et hiérarchiser. Le Pôle Santé République pilote des groupes de travail réunissant toutes les ressources internes impliquées dans leur EDS et leurs partenaires externes afin d'optimiser le flux de traitement des données. Cet établissement lance systématiquement une étude de faisabilité préalable au lancement de chaque projet.

UN SCORE D'ÉVALUATION

À partir des différentes dimensions explorées dans le questionnaire administré aux établissements pilotes, nous avons sélectionné 27 paramètres discriminants. Ces paramètres ont été répartis en 4 dimensions (dont Usage des données, Moyens alloués et Gouvernance des données) avec tests puis associations de pondérations aux différentes modalités des 27 paramètres. Nous avons ainsi élaboré et testé un score par dimension et un score global tenant compte des caractéristiques intrinsèques de chaque type d'établissement. Ces scores pourront permettre, à partir d'un questionnaire autoadministré sous forme de chatbot par exemple, de suivre annuellement l'évolution de son établissement dans la maîtrise et l'acquisition des compétences et des moyens pour favoriser l'utilisation des données de soins dans une démarche de recherche en épidémiologie et de recherche en Santé publique.

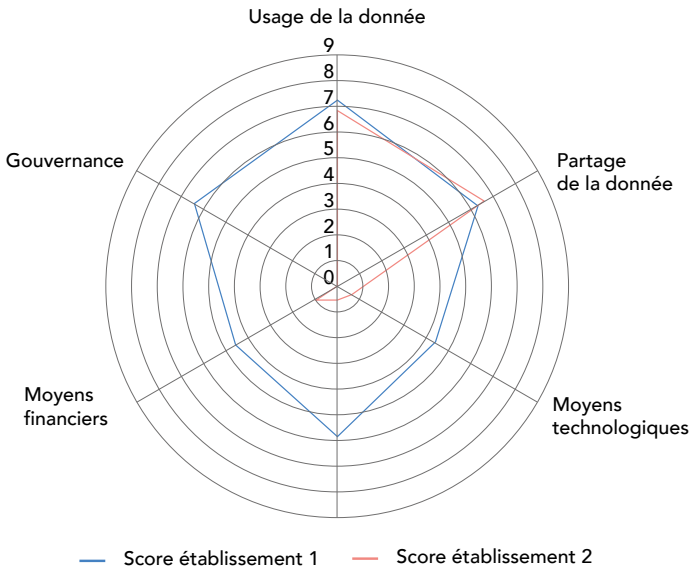


Figure 3 : exemple de scores entre 2 établissements

PARTIE 3

LA VILLE, LES PATIENTS ET LES DONNÉES

Les professionnels de santé de la ville (médecins, pharmaciens, kinés, infirmiers, laboratoires d'analyse médicale, etc.) voient depuis quelques années une profonde évolution de leurs missions. Le décloisonnement « ville-hôpital » demeure le défi majeur pour notre système de santé. Alors que l'on assiste au développement des maladies chroniques, au vieillissement de la population, au renforcement de certaines inégalités d'accès aux soins, notamment territoriales, il devient de plus en plus nécessaire de pouvoir accompagner les patients dans un parcours de soins global, coordonné, incluant la ville et l'hôpital. Une démarche de décentralisation des soins et de la prise en charge des patients est fortement engagée afin de réduire les inégalités territoriales. Pour lutter contre la désertification médicale, des réflexions sont en cours pour attribuer des missions complémentaires aux infirmiers, pharmaciens, dentistes et soulager les médecins généralistes. Les réseaux de soins régionaux se structurent, les programmes RAAC (Récupération Accélérée Après Chirurgie) offrent une meilleure prise en charge du patient, et un gain en matière de lits d'hôpitaux/rationalisation des dépenses de santé.

Comme pour l'hôpital, la structuration et la capacité de centraliser les données de ville sont assez efficaces pour les données impactant directement la prise en charge financière par la collectivité, que ce soient les données issues de la télétransmission des actes de ville ou de dispensation des traitements en ville. En revanche, pour ce qui concerne les données médicales, la médecine de ville nécessitera aussi une optimisation de la collecte et du partage des données des patients pour la réalisation d'études épidémiologiques ou santé publique avec pour ambition première, l'amélioration du parcours de santé des patients.

Une stratégie politique dédiée aux données de ville devrait également être mise en œuvre pour sensibiliser les professionnels de santé sur l'importance de partager les données issues de leurs patients, dans un cadre réglementaire adapté et de façon sécurisée et standardisée. Les professionnels de santé doivent être sensibilisés et accompagnés dans la démarche de collecte et de partage des données pour contribuer à la recherche, dans un intérêt de santé publique ou encore dans un intérêt de meilleure prise en charge de leurs patients.

Ségur du numérique en santé pour les médecins de ville[19]

À partir de 2023, l'état va subventionner via le Ségur du numérique les médecins de ville afin qu'ils puissent s'équiper avec des logiciels compatibles avec l'espace Patient *Mon Espace Santé*. Parmi les avantages de passer sur ces outils validés : les échanges sont facilités et sécurisés avec l'hôpital, utiliser la e-prescription ou encore structurer la saisie des diagnostics et des médicaments.

Tout comme l'accompagnement des médecins lors de l'informatisation de l'exercice médical et la télétransmission, on pourrait imaginer la mise en place d'un forfait «Données de vie réelle» aux ROSP (rémunération sur objectifs de santé publique) actuelles pour les médecins, et étendu à tous les professionnels de santé impliqués dans la prise en charge et le suivi des patients pour aider et motiver les professionnels de santé à collecter et transmettre les données de leurs patients. On sait qu'environ 80 % des médecins généralistes, plus de 85 % des infirmier(e)s et près de 100 % des pharmaciens utilisent la télétransmission contre environ 55 % des médecins spécialistes, et sont donc équipés pour transmettre des données. En revanche, cela nécessite de disposer d'une codification implantée dans tous les logiciels et d'organiser la remontée des données en temps réel. La mise en place d'un tel forfait semble également en droite ligne de l'objectif initial des ROSP qui est de contribuer à faire évaluer les pratiques pour atteindre les objectifs de santé et pas seulement dans les maladies chroniques. La mise en œuvre d'un retour systématique de données vers le praticien de ville ou vers l'hôpital, sur le parcours des patients pris en charge, pourrait grandement favoriser l'intérêt des professionnels à la collecte et à l'analyse des données de santé.

Côté recherche, il existe déjà en France et en Europe des observatoires de ville permettant de fournir à tous les acteurs de santé, y compris les services de l'État (ANSM, CEPS, NICE, NHS) et l'EMA des analyses médicalisées, des études prélancement, des modélisations de patients souffrant de maladies rares, des observatoires épidémiologiques, des études académiques, en lien avec la pratique de ville, montrant l'intérêt de suivre l'histoire du patient avec une base de données médicalisée issue de la ville. De nombreuses recherches s'appuyant sur des données de vie réelles de Primary Care ont fait l'objet de publications : plus de 3000 publications CPRD[20,21] et plus de 2000 THIN[22].

Focus sur les initiatives menées dans les maisons de santé

Ce modèle de groupe de structures de santé de proximité initié en Suède a également séduit les cliniques privées. Les maisons de santé ont un objectif de suivi populationnel et pluridisciplinaire au niveau local.

Comme pour les cabinets libéraux, le Ségur du Numérique a approuvé des outils adaptés aux maisons de santé qui vont faciliter l'interopérabilité des systèmes d'information notamment entre la ville et l'hôpital.

Plusieurs initiatives[23] menées dans les maisons de santé montrent que le recrutement d'un infirmier, d'un assistant médico-administratif permet d'accroître le nombre de patients vus en consultation et d'améliorer la collecte des données. Ces emplois sont financés par l'augmentation de la patientèle traitée. Réussite éprouvée pour plusieurs spécialités (médecine générale, cardiologie, chirurgie dentaire, ophtalmologie).

Ramsay Santé et Elsan ont déjà lancé des structures de ce type. L'offre comprend des médecins généralistes, des spécialistes, du personnel paramédical et un service de téléconsultation.

L'offre peut également compter un pharmacien ou un partenariat avec une officine de ville.

Aussi, un volet médical de synthèse sous le modèle de La lettre de liaison (LL) visant à améliorer la coordination entre les professionnels de santé lorsqu'un patient sort d'un service hospitalier. La transmission des données est capitale pour assurer et sécuriser la prise en charge en aval lorsque le patient rentre à domicile ou est transféré dans un autre service ou établissement (établissement de santé (ES). Dans ce cadre, un livret thérapeutique commun pourrait diminuer les risques pour le patient (rupture éventuelle dans la continuité de prise en charge médicamenteuse en lien avec le changement de spécialité ; iatrogénie médicamenteuse par surdosage, oubli, redondance).

De la ville vers hôpital : l'inobservance constatée par les praticiens de ville, en premier lieu les pharmaciens, nécessiterait un partage de données avec leurs homologues hospitaliers par différents supports : DMP, Dossier pharmaceutique toute rupture de parcours. Un projet « Pharm'Observance PACA » (https://www.urps-pharmaciens-paca.fr/le-projet/) est en cours porté par l'URPS Pharmacien PACA afin d'améliorer l'observance des traitements prescrits pour les maladies chroniques en coordination avec l'ensemble des acteurs.

La pharmacie prendra aussi une place prépondérante dans le parcours de soin de ville comme par exemple les entretiens pharmaceutiques et le suivi du parcours de soins du patient. Compte tenu de la fréquentation des pharmacies et des données recueillies, celle-ci pourrait devenir un véritable hub de data pour les patients.

Le patient, au centre des enjeux de la recherche en santé publique, pourra aussi être acteur du partage de ses données pour le bénéfice de la communauté. Une offre considérable d'objets connectés et d'applications mobiles sera à la disposition du patient pour l'assister dans la prise en charge de sa pathologie. La HAS a notamment publié un guide d'évaluation des dispositifs médicaux faisant appel à l'intelligence artificielle en 2019 pour assurer un réel bénéfice pour le patient.

Que ce soit pour une prise en charge des patients en ville ou à l'hôpital, les questionnaires de type PREMs et PROMs vont également contribuer à apporter la vision des patients sur la qualité des soins. Il ne fait aucun doute que ces outils seront riches d'exploitation pour la recherche en santé publique et l'épidémiologie.

PARTIE 4

RECOMMANDATIONS POUR DÉCLOISONNER L'USAGE DES DONNÉES DE SANTÉ

Nous avons regroupé les recommandations en 2 chapitres fonction des destinataires des recommandations.

Le premier paragraphe concerne les établissements de santé, organisé sous la forme de 4 groupes de recommandations : 1) Mise en place d'un plan stratégique de réutilisation des données de santé en lien avec les missions de l'hôpital ; 2) Mise en œuvre d'un plan d'équipement pour faciliter la ré-exploitation des données de soins produites ; 3) Se mettre en conditions optimales afin de développer des partenariats externes ; 4) Mettre en œuvre une organisation adaptée aux ambitions et aux expertises de l'établissement de santé intégrant les partenaires adéquats.

Le second chapitre est à destination des autorités de santé et des décideurs (AD), organisés en 3 groupes de recommandations : 1) Promouvoir une production de données de qualité utilisables pour d'autres recherches que celles d'établissement de Santé ; 2) Mesurer l'évolution des usages et la valeur produite par ces données pour tous les acteurs ; 3) Créer un cadre de confiance collectif (public/privé) et promouvoir l'information et la transparence collectivement sur l'usage des données de santé pour l'intérêt collectif

À DESTINATION DES ÉTABLISSEMENTS DE SANTÉ

Stratégie et planification intégrée	→ Se doter d'un plan stratégique de réutilisation des données de santé en lien avec les missions de l'hôpital (pourquoi, readiness, moyens, structuration, planification…)
ET 1	Mettre au niveau du comité de direction des établissements de santé une représentation liée à la réutilisation des données
ET 2	Définir des ambitions/objectifs (internes, externes) d'exploitation de données et des indicateurs de valeur et soutenabilité plan
ET 3	Avoir une approche graduelle dans la mise en œuvre des moyens et compétences en fonction de votre stratégie
ET 4	Identifier les cibles d'exploitation possibles et la meilleure manière pour atteindre cette cible (en propre, raccordement à un réseau de données existant, exploitation *via* des sous-traitants, registres/cohortes)

Technologie et data	→ Se doter d'un plan d'équipement pour faciliter la ré-exploitation des données de soins produite
ET 5	Développer une approche de standardisation des données pour les usages secondaires
ET 6	Se doter d'outils de réutilisation des données de votre établissement de santé – conformes aux exigences techniques de sécurité recommandées
ET 7	Intégrer dans les nouveaux systèmes digitaux de production de soin des moyens facilités pour la ré-exploitation des données (télémédecine, dossiers médicaux, etc.)
ET 8	Avoir une stratégie ciblée sur la qualité des données produites

Valorisation et utilisation	→ Se mettre en conditions (readiness) optimales afin de développer des partenariats externes
ET 9	Partager de manière transparente, publiquement, les services, les données disponibles, les savoir-faire, les moyens humains, les conditions et modalités d'accès, les délais pour déployer des projets par type de projet
ET 10	Réduire les délais de contractualisation : bien scinder les discussions relatives à l'investissement d'équipement versus de valorisation des données (mieux connaître ses données, clarifier l'apport externe pour mise à niveau de la donnée, approche règlementaire rapide, modèles par type d'acteur/type de projet, gestion de la PI, sous-traitance, faisabilité plus rapide)
ET 11	Définir des modèles économiques a priori prenant en compte la prise de risque, la durée, l'amortissement de l'infrastructure et les investissements futurs
ET 12	Utiliser des mécaniques de mitigation des risques d'investissements externes (appels à projets) pour accompagner les projets (au stade de faisabilité, ou si les intérêts sont partagés)
ET 13	Évaluer et observer ses progrès, partager les expériences entre pairs pour s'améliorer

Organisation et partage	→ Mettre en œuvre une organisation adaptée aux ambitions et aux expertises de l'établissement de santé intégrant les partenaires adéquats
ET 14	Intégrer des compétences médicales (comprendre les données : compréhension du contexte de collecte primaire et interprétation des données), sciences des données, gestion de projet et valorisation pour développer les infrastructures de données et les projets
ET 15	Promouvoir des métiers et des organisations au sein des établissements de santé qui permettent une efficience plus grande de la captation des données pour la recherche observationnelle (capitalisation)
ET 16	Mettre en œuvre une approche structurelle d'un côté, et projet de l'autre
ET 17	Créer des structures support au sein des établissements de santé pour accompagner la mise en œuvre des projets, distincte des équipes médicales ou recherche conventionnelles
ET 18	Les équipes internes peuvent se concentrer sur les moyens nécessaires à l'organisation, l'accompagnement des projets, la stratégie et la valorisation. Pour le reste, ne pas hésiter à s'adjoindre des compétences externes (data science, analyses, rédaction médicale, etc.).
ET 19	Intégrer les exigences liées à la sécurité des données dans le plan de gouvernance global mis en place autour de la donnée.
ET 20	Partager les expériences entre établissements de santé et avec les leaders et les autorités de santé (HAS en particulier dans un objectif de communication/visibilité)

À DESTINATION DES AUTORITÉS DE SANTÉ (NOTAMMENT HAS) ET DÉCIDEURS

Reconnaître les producteurs de données	→ Promouvoir une production de données de qualité utilisables pour d'autres recherches que celles de l'établissement de santé
AD 1	Reconnaître le travail de mise en qualité et valorisation des données fait généralement par l'exploitant de données (public/privé) en complément du travail du producteur primaire (coresponsabilité de traitement, propriété intellectuelle sui generis)
AD 2	Adapter les modèles de financement public pour les producteurs de données utilisables dans le cadre de l'évaluation des produits de santé et de santé publique. Par exemple, une reconnaissance des indicateurs RNIPH pour tous les producteurs de données participant aux études. Sur un financement pour une exploitation de données une cote part de 50 % pourrait être versée au producteur de données et 50 % pour à l'exploitant. Un mécanisme de type ROSP pourrait être aussi adapté pour les producteurs de ville.

Piloter et prioriser	→ Mesurer l'évolution des usages et la valeur produite grâce à ces données pour tous les acteurs
AD 3	Mesurer la mise à l'échelle et l'amplification de la réutilisation pour/par tous les acteurs par le biais d'un observatoire national
AD 4	Identifier les enjeux de santé publique prioritaires nécessitant de cibler certaines actions publiques de réutilisation des données
AD 5	Evaluer les politiques publiques d'investissements en développant des indicateurs d'impact objectifs d'intérêt général, ex : % de données re-utilisées versus régénérées, production scientifique
AD 6	Capitaliser sur les Co-investir (renforcer) aussi dans les réseaux/bases de données qui produisent déjà de la connaissance
AD 7	Identifier les infrastructures de données à visée de santé publique en capacité de produire des données de qualité, soutenable pour le système de santé (soutien commun de plateformes) et attractives pour les investisseurs (transparentes, pilotées), des mécaniques de co-investissement (public/privé)
AD 8	Mettre en œuvre une politique ambitieuse sectorielle en termes de standardisation des données générées (approches données minimales, values établissements de santé, terminologies, modèles de données, etc.…)

Établir des cadres de confiance	→ Créer un cadre de confiance collectif (public/privé) et promouvoir l'information/ transparence ensemble sur l'usage des données de santé pour l'intérêt collectif
AD 9	Labelliser des circuits de partage et d'exploitation de données de confiance (public ou privé) thématiques (santé publique, par maladie, etc.), certifiable, auditable, et couvrant les dimensions technico-réglementaires et leur usage scientifique
AD 10	Former/communiquer/sensibiliser sur le bien-fondé de la réutilisation des données, les exigences de sécurité, pour la production de connaissances médicales et scientifiques
AD 11	Produire des méthodologies de référence CNIL liées à l'exploitation des produites par les acteurs du soin (hospitaliers) pour la santé publique valables pour tous les acteurs exploitants
AD 12	Partage systématique des données d'activité, via l'envoi des données pour le PMSI, qui pourrait être accompagnée d'un retour systématique de données issues de l'activité libérale disponibles dans le SNDS.

CONCLUSION

La réutilisation des données issues du soin pour la recherche et la santé publique n'est pas une activité triviale pour un établissement de santé. Elle nécessite un engagement fort de l'établissement car il mobilisera nécessairement des ressources rares. Bien que les engagements politiques soient toujours plus importants dans le domaine des données de santé, il n'en reste pas moins que les recherches en santé publique nécessitent des données de qualité, représentatives de la prise en charge, donc, provenant de plusieurs établissements de santé en même temps. Comme le souligne la HAS dans son récent rapport[24] sur les entrepôts de données de santé hospitaliers en France, l'état actuel d'implémentation de ces approches est trop hétérogène pour sérieusement envisager l'usage de ces données pour l'évaluation ou des recherches à l'échelle nationale. Dans ce même rapport, la HAS propose des recommandations complémentaires à celles proposées dans ce Position Paper.

Le travail collaboratif mené par le HDI sur la meilleure réutilisation des données hospitalières issues de la production des soins pour la santé publique dresse un bilan mitigé du sujet à la maille hospitalière. L'utilisation des données de soins pour des usages dépassant la mesure de l'activité est encore trop hétérogène. L'approche multicentrique semble encore trop complexe à aborder. La plupart des projets externes (d'intérêt externe à l'établissement) se heurtent à des temps de mise en œuvre trop longs et bien souvent la question de la faisabilité. Enfin, la sécurisation des entrepôts, point hautement stratégique et dont le cadre a bien été défini par la CNIL, reste un sujet technologique complexe, multidimensionnel, requérant des expertises avancées, mais qui doit impérativement rester sous contrôle des établissements fournisseurs de données. Il s'agit ici d'opérer une profonde transformation de la manière dont tout acteur de l'offre de soins participe à la production de données de santé de qualité pour la santé publique. Chaque maille doit s'engager en toute connaissance de cause. Partager son expérience avec ses pairs. Dans le même temps, différentes modalités de financement doivent pouvoir s'intégrer dans le temps long, et chaque acteur doit pouvoir, sur les mêmes données, capitaliser pour produire les connaissances nécessaires à chacun.

Pour épauler le système dans ce travail, nous avons proposé des indicateurs préliminaires dans un outil, testé auprès de premiers établissements, afin de mesurer et de suivre l'évolution de la maturité de la démarche afin de permettre à chaque établissement de suivre et progresser dans l'identification de ses forces et de ses faiblesses pour adapter son organisation pour la gestion des données de soins pour la recherche. Dans la continuité de ce travail, un observatoire sur l'usage des données de soins pour la recherche permettra à chaque établissement, en accord avec sa politique et ses objectifs, de suivre et progressivement s'impliquer dans des projets avec données de soins partagées. Avec le soutien des groupements hospitaliers volontaires et une politique de soutien ambitieuse, de nouveaux objectifs de projets de recherche en épidémiologie et de recherche en Santé publique s'appuyant sur un partage maîtrisé de données, seront source de nombreuses nouvelles connaissances dont la finalité restera une meilleure prise en charge des patients.

Pour accompagner la démarche, nous avons proposé 20 recommandations complémentaires, visant à promouvoir, au-delà de la démarche engagée par le HDH, une approche cohérente, stratégique, et étagée au sein des établissements de santé afin de promouvoir une montée en charge incrémentale, mais systémique d'une approche data. Enfin, nous proposons 12 recommandations à destination des pouvoirs publics afin d'aider chaque producteur de données, offreur de soins, à s'inscrire dans une démarche plus multicentrique, ciblée vers des priorités de santé publique communes.

Une politique volontaire de partage et de valorisation des données de soins au bénéfice de la collectivité s'appuie sur soutien institutionnel volontaire, passant par exemple par un accompagnement spécifique et une meilleure reconnaissance des travaux menés sur ces données.

ANNEXE 1

REMERCIEMENTS

Le groupe de travail tient à remercier les centres ayant participé au test du questionnaire et qui ont validé les recommandations proposées.

Établissement	Contact
CHU de Reims	Pr. Vincent VUIBLET, directeur de l'institut d'intelligence artificielle en santé Reims Champagne-Ardenne (I2AS)
Centre hospitalier de Valenciennes	Laurie FERRET, pharmacie et unité de recherche clinique Icham SÉFION, directeur des systèmes d'information
Institut Bergonié	Sylvie CASSAUBA, directrice des systèmes d'information Julien-Aymeric SIMONNET, directeur Données et Santé Numérique
Clinique Pôle Santé République, ELSAN	Frédéric JOURDAN, directeur administratif et financier
Centre de cancérologie Les Dentellières, ELSAN	Emmanuelle MAZER, directrice
Polyclinique Bordeaux Nord Aquitaine	Dr Nadine DOHOLLOU, oncologue sénologue Delphine BRUNIE, attachée de recherche clinique Marie LASCAUD, administratrice société radiothérapeute
Centre hospitalier de la Côte Basque	Dr Anne CHAMBON, pharmacien, coordonnateur de la recherche clinique du CHCB
Centre hospitalier départemental de Vendée	Valérie DESROYS DU ROURE, biologiste, responsable de l'URC, référente investigation Philippe FEIGEL, chef de service, département d'information médicale de territoire, unité de recherche clinique
Centre hospitalier de Troyes	Dr Stéphane SANCHEZ, MCU-PH associé, épidémiologie, santé publique et prévention, responsable unité de recherche clinique et recherche en soins, responsable adjoint du pôle territorial santé publique et performance
Centre Léon Bérard	David PÉROL, directeur de la recherche clinique et de l'innovation Frédéric GOMEZ, directeur de l'information médicale
Centre hospitalier de Villefranche-sur-Saône	Marie-Pierre BONGIOVANNI VERGEZ, direction générale Nasser AMANI, direction des services numériques Lionel FALCHERO, pneumologie Adeline ANDRÉ, département de l'information médicale

Ainsi que les experts auditionnés :
Alexis HECHT, Charlotte FRABOULET, Marc CUGGIA, Thomas SÉJOURNÉ, Anca PETRE.

ANNEXE 2

TÉMOIGNAGES DE DEUX ÉTABLISSEMENTS SOLLICITÉS

Centre hospitalier de Valencienne
Laurie FERRET, responsable de la pharmacie hospitalière
Icham SEFION, direction des systèmes d'information

1/ Aujourd'hui dans votre établissement, estimez-vous que les données collectées dans vos systèmes d'information sont facilement exploitables à visée de santé publique ou d'étude épidémiologique ?

L'exploitabilité des données dépend de nombreux facteurs, tels que leur accessibilité, leur nature, le nombre de sources, etc. La facilité de mise en place d'une étude dépend aussi de questions réglementaires, contractuelles, et d'un cheminement institutionnel. L'ensemble du processus est un facteur de succès pour les études sur données du système d'information.

Au CHV, un certain nombre de projets nécessitant des données extraites du DPI sont en cours. Pour certains d'entre eux, les difficultés rencontrées nous ont permis d'entrer dans une démarche d'amélioration du processus global, à la fois sur le plan technique et l'accompagnement des porteurs de projets.

2/ Quels sont les grands challenges auxquels vous faites face et les perspectives de développement dans votre établissement sur ce sujet ?

De plus en plus de projets innovants sont initiés au CHV en exploitant nos données de santé. La majorité des projets intègre le développement ainsi que l'évaluation d'algorithmes d'intelligence artificielle. Ils sont fréquemment menés en partenariat avec des entreprises privées spécialisées.

Ces projets nécessitent en interne un support sur les plans technique/informatique, méthodologique et juridique, et suivent le circuit institutionnel adéquat selon le projet. Une coordination efficace entre les différents acteurs et une optimisation de leur organisation autour de ce type de projets sont des enjeux importants pour fluidifier le processus et ainsi mieux accompagner les porteurs de projet.

3/ Pouvez-vous nous partager un projet piloté dans votre établissement sur ce sujet ? Un projet « totem » ?

Un comité IA-innovation a été créé en 2022. Il rassemble les différentes cellules support de l'établissement ainsi que des porteurs de projets de recherche intégrant de l'intelligence artificielle. Il assure un suivi global des différents projets afin de les mener à terme.

Le comité se réunit régulièrement afin d'échanger sur les problématiques rencontrées par les porteurs de projets et réfléchir à des solutions à l'échelle de chaque projet, mais aussi à l'échelle de l'établissement. Pour faciliter les démarches, améliorer l'accompagnement par les cellules support et faciliter la mise en œuvre des projets de recherche.

Clinique Pôle Santé République de Clermont-Ferrand – groupe Elsan
Frédéric JOURDAN, directeur administratif et financier
Olivier FAURE, responsable technique et logistique

1/ Aujourd'hui dans votre établissement, estimez-vous que les données collectées dans vos systèmes d'information sont facilement exploitables à visée de santé publique ou d'étude épidémiologique ?

La progression ces dernières années de l'informatisation et de la numérisation des échanges a permis de réelles avancées, même si certains sujets restent à traiter (ex. : hémovigilance) ou ont été pris en compte récemment (ex. : outil digital RCP en 2020).

À ce jour, nos systèmes d'information hospitaliers (SIH) ont une colonne vertébrale de datas solides et relativement facilement exploitables. Cependant, la saisie des données de santé reste hétérogène, ce qui implique pour chaque projet une étude de faisabilité en fonction du cahier des charges.

2/ Quels sont les grands challenges auxquels vous faites face et les perspectives de développement dans votre établissement sur ce sujet ?

Il s'agit surtout d'un changement de paradigme dans la manière dont les médecins et soignants abordent la saisie dans le dossier patient. Dans un contexte de pénurie médicale, c'est un sujet particulièrement sensible.

L'harmonisation et la structuration des datas font partie des principaux enjeux. En effet, les données sont d'autant mieux réutilisables qu'elles sont structurées au moment de la saisie dans le dossier patient, ce qui n'est pas toujours le cas. L'information peut donc être présente, mais difficilement exploitable. L'évolution de nos SIH doit le prendre en compte et s'adapter aux contraintes des utilisateurs médecins et soignants. Nous faisons un travail conjoint important avec la communauté médicale, la DSI et nos éditeurs afin d'accompagner l'harmonisation des données recueillies dans le cadre des soins, que ce soit en termes de structure, de nomenclature et de complétude. Sur ce dernier point, l'interopérabilité et la collaboration entre les professionnels de santé et leurs éditeurs sont de vrais challenges.

En complément, ces sujets très complexes sont soutenus par une politique nationale de qualité et de conformité des SIH, notamment dans le cadre des programmes d'accompagnement du numérique (Ségur…). Le Groupe ELSAN s'est également investi dans une démarche de meilleure utilisation et valorisation des données de santé. Aujourd'hui, le niveau d'avancée de nos systèmes d'information constitue un outil d'aide à la décision stratégique et un support à la construction de projets médicaux à différents niveaux (interspécialités, interétablissements…).

Au-delà, nos SIH constituent une base de données de vie réelle avec une complémentarité aux études cliniques qui ouvrent des perspectives aux praticiens et aux promoteurs pour leurs projets de recherches.

3/ Pouvez-vous nous partager un projet piloté dans votre établissement sur ce sujet ? Un projet « totem » ?

Notre établissement offre une diversité importante de soins et réunit une large communauté médicale, ce qui engendre une multiplication des besoins métiers qui doivent être adressés par le même SIH.

Le niveau de maturité croissant de notre SIH nous permet de mener de plus en plus d'études. À titre d'exemple, nous avons pu récemment : accompagner des projets médicaux portés par nos praticiens (ex. : suivi du virage ambulatoire du service de cardiologie interventionnelle…) ; évaluer le bénéfice de nouveaux parcours de soins pour les patients (impact de la coordination des soins en oncologie sur le nombre d'hospitalisations en urgence, les effets indésirables et le coût de la prise en charge ; faciliter l'intégration des innovations (étude d'impact de l'intrabeam, radiothérapie peropératoire, sur le parcours patient : hospitalisations, transports sanitaires, taux de rechute à 5 ans…)

Plus récemment, nous avons été retenus parmi les 10 sites pilotes du programme OncoDataHub (ODH) lancé par UNICANCER et le laboratoire Roche pour créer une plateforme de référence de données de vie réelle en oncologie. Ce projet nous a permis d'avancer dans notre connaissance des forces et limites de nos données de santé et surtout nous pousse à étendre le périmètre et la qualité de notre SIH et des données qui s'y trouvent.

TABLE DES RÉFÉRENCES

1. Haut Conseil de la Santé PubliqueHCSP. Registres et données de santé : Utilité et perspectives en santé publique [Internet]. Rapport de l'HCSP. Paris : Haut Conseil de la Santé publique ; 2021 Sept. Available from: https://www.hcsp.fr/Explore.cgi/avisrapportsdomaine?clefr=1126

2. Bégaud, B, Polton D, Von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé : L'exemple du médicament [Internet]. 2017 Mai. Available from: http://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf

3. Health Data Hub. Un AAP de 50 M€ pour la constitution d'Entrepôts de données de santé hospitaliers [Internet]. Health Data Hub. 2022. Available from: https://www.health-data-hub.fr/actualites/aap-entrepots-donnees-de-sante-hospitaliers

4. Research C for DE and. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products [Internet]. Food and Drug Administration; 2021. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

5. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574–8.

6. Data Standardization – OHDSI [Internet]. [cited 2022 Dec 5]. Available from: https://www.ohdsi.org/data-standardization/

7. FHIR v4.3.0 [Internet]. Available from: https://hl7.org/fhir/

8. Système national des données de santé – Référentiel de sécurité - Guide pédagogique du référentiel de sécurité.

9. eHOP : Enovacom alimente le premier entrepôt de données d'Europe [Internet]. Enovacom. 2022. Available from: https://www.enovacom.fr/thematiques/echanges-de-donnees/ehop-un-partenariat-innovant-entre-enovacom-et-le-chu-de-rennes-autour-dun-entrepot-de-donnees-de-sante

10. Institut Imagine, guérir les maladies génétiques [Internet]. Institut Imagine. 2020. Available from: https://www.institutimagine.org/fr/nicolas-garcelon-757

11. Qu'est-ce que l'Entrepôt de Données de Santé (EDS) ? [Internet]. Direction de la Recherche clinique et de l'Innovation de l'AP-HP. 2016. Available from: https://recherche.aphp.fr/eds/definition/

12. Rapport 2021 : Observatoire des signalements d'incidents de sécurité des systèmes d'information pour le secteur santé [Internet]. Available from: https://www.vie-publique.fr/sites/default/files/rapport/pdf/284968.pdf

13. Commission Nationale de l'Informatique et des Libertés. CNIL – Délibération n° 2021-118 du 7 octobre 2021 portant adoption d'un référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé – Légifrance [Internet]. 2021–118 Oct 24, 2021. Available from: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044239566

14. genomEditor4All. A conversation on Federated Learning [Internet]. GenoMed4All. 2022. Available from: https://genomed4all.eu/2022/07/06/a-conversation-on-federated-learning-part-1/

15. St. Michael's Hospital Toronto – About us [Internet]. LKS-CHART. Available from: https://www.chartdatascience.ca/about-us

16. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Intern Med. 2021 Aug 1;181 (8):1065.

17. Antoniou T, Mamdani M. Évaluation des solutions fondées sur l'apprentissage machine en santé. Can Med Assoc J. 2021 Nov 8;193(44):E1720 –4.

18. DREES. Études et Résultats. septembre 2022. n° 1242 [Internet]. Available from: https://drees.solidarites-sante.gouv.fr/sites/default/files/2022-09/ER1242-EMB.pdf

19. Ministère de la Santé et de la Prévention. Le Ségur du numérique en santé [Internet]. 2022. Available from: https://esante.gouv.fr/segur

20. Bibliography | CPRD [Internet]. Available from: https://www.cprd.com/bibliography

21. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile : Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015 Jun;44(3):827–36.

22. Resources Hub | THIN Data [Internet]. Available from: https://www.the-health-improvement-network.com/resources-hub

23. Rapport du HCAAM – Organisation des Soins de proximité : Garantir l'accès de tous à des soins de qualité [Internet]. 2022 Sept. Available from: https://www.strategie.gouv.fr/sites/strategie.gouv.fr/files/atoms/files/rapport_hcaam_organisation_soins_proximite.pdf

24. Haute Autorité de Santé – Entrepôts de données de santé hospitaliers : la HAS publie un panorama inédit en France [Internet]. Available from: https://www.has-sante.fr/jcms/p_3386076/fr/entrepots-de-donnees-de-sante-hospitaliers-la-has-publie-un-panorama-inedit-en-france

AUTRES LIENS UTILES

Réglementaire :

https://www.cnil.fr/fr/la-cnil-adopte-un-referentiel-sur-les-entrepots-de-donnees-de-sante

Remboursement :

https://www.has-sante.fr/jcms/p_3318028/fr/grille-descriptive-des-fonctionnalites-des-dispositifs-medicaux-embarquant-un-systeme-avec-apprentissage-automatique-intelligence-artificielle

https://www.has-sante.fr/upload/docs/application/pdf/2021-06/guide_etude_en_vie_reelle_medicaments__dm.pdf

Politique publique d'investissement :

https://www.health-data-hub.fr/actualites/ami-csf-its

https://www.usine-digitale.fr/article/le-health-data-hub-a-ete-choisi-pour-mener-le-futur-espace-europeen-des-donnees-de-sante.N2028137

Nouvelles technologies :

https://medcitynews.com/2022/09/owkins-ai-solutions-for-breast-and-colorectal-cancer-receive-european-approval/

https://octopize-md.com/en/

HEALTHCARE DATA INSTITUTE

HEALTHCARE DATA INSTITUTE
Chez RCA Factory, 39, rue d'Aboukir, 75002 Paris

@HCDataInstitute | **healthcaredatainstitute.com**

CONTACT
office@healthcaredatainstitute.com | +33 (0)1 42 21 19 59