

CONTRAT DU COMITÉ STRATÉGIQUE DE LA FILIÈRE INDUSTRIES ET TECHNOLOGIES DE SANTÉ

2023-2026



**Rapport du groupe de travail – Exploiter et valoriser les
données de santé en vie réelle**

**De la donnée à sa valorisation : Améliorer la qualité
des données de santé en vie réelle pour un usage
secondaire pertinent**

Mars 2026

Remerciements

Le groupe de travail “ Exploiter et valoriser les données de santé en vie réelle ” tient à remercier l’ensemble des personnes consultées pour la réalisation des travaux du groupe de travail :

- Emilie COURTOIS, Jérôme DAUMAS et Alexis HECHT, pour le HDH (juillet 2024) : Entretien avec le CSF IA et Santé, destiné à comprendre les référentiels utilisés par le HDH dans l’identification des bases d’intérêt, et l’implication du HDH dans l’initiative QUANTUM.
- Lauriane BROTTIER (septembre 2024) : Partage des conclusions de sa thèse intitulée « Comment les données de vie réelle sont-elles appréhendées par la HAS lors des soumissions de dossiers d’oncologie à la CT ? ».
- Dr Guillaume MOULIS, pour le registre Carmen (septembre 2024) ; Pr Pascal AUQUIER, et Dr Clémence Tabélé pour le registre FranceCoag (octobre 2024) ; Pr Pierre-Antoine GOURRAUD, et Etienne BENDJEBBAR, pour la Clinique des Données du CHU de Nantes (octobre 2024) ; Samantha PASDELOUP, pour ELSAN (octobre 2024) ; Géry PRUVOST, pour Lifen (septembre 2024) : Entretiens avec le CSF IA et Santé, destinés à définir de premières pistes de bonnes pratiques et identifier les challenges rencontrés par les producteurs de données.
- Stéphanie LECLERC, Judith FERNANDEZ et Camille THOMASSIN, pour la HAS (septembre 2024) : Entretien avec le CSF IA et Santé, autour des sujets d’acceptabilité et de recevabilité des études construites sur données secondaires.
- Les cohortes et registres FREGAT (CHU Lille), ESME (Unicancer), OFSEP (HCL), ODH (Unicancer), NS-Park (AP-HP), LUCC (Lifen), Constances (Inserm), FranceCoag (AP-HM), REALYSA (Lysarc), Sacha (Institut Gustave Roussy), SMA (AP-HP), REA-REZO (HCL), OMIN (CHU Nantes), Uro-CCR (CHU Bordeaux) : Réponse à l’enquête administrée en ligne entre mai et juillet 2025.
- Les entrepôts de données de santé hospitaliers Nantes (CHU Nantes), GCS HUGO (CHU Angers), ELSAN, Lille (CHU Lille) : Réponse à l’enquête administrée en ligne entre mai et juillet 2025.

De la donnée à sa valorisation

Améliorer la qualité
des données de santé en vie réelle
pour un usage secondaire pertinent



Recommandations du Groupe de Travail

Axe stratégique	Actions clés (incluant registres/cohortes & EDSH)	Objectif & Bénéfice	Horizon	Contributeurs
Indicateurs qualité	Définir un set d'indicateurs essentiels (complétude, suivi, représentativité). <i>Pour Registres/cohortes : publier les indicateurs clés (open data, accord de confidentialité simple).</i> <i>EDSH : transparence des métriques par type de données (PMSI, biologie, comptes rendus, imagerie).</i>	Uniformiser les pratiques et faciliter les décisions	Court terme	HAS, HDH, Producteurs de données
Transformation & structuration des données	Développer l'informatisation à la source, améliorer l'interopérabilité, et accompagner la structuration via NLP/IA. <i>EDSH : cadre qualité spécifique, harmonisation des comptes rendus.</i>	Optimiser la qualité des données et leur exploitation	Court terme	HDH, CHU, Sociétés savantes, Entreprises de la Health Tech, AFCROs
	Spécifier un référentiel de variables minimales normalisées.	Assurer compatibilité et échanges automatisés	Court terme	HDH, ARIIS, Ministère Santé, INSERM
Plateforme nationale d'information	Capitaliser sur les initiatives en cours (type FReSH) pour créer un catalogue unique recensant registres, cohortes et EDSH avec métadonnées, labels QUANTUM et moteur de recherche par pathologie.	Améliorer la transparence et augmenter l'attractivité des bases de données existantes françaises	Moyen terme	HDH, CNIL, ARIIS, AFCROs, Entreprises des Produits de Santé
Charte nationale de transparence	Définir des niveaux d'accès (open data, accords de confidentialité, contrat) et délais standards aux bases.	Clarifier les règles de partage et accélérer les projets	Court terme	HAS, Ministère Santé, Producteurs de données
Formation & professionnalisation	Créer un cursus national certifiant (incluant notamment les sujets data management, RGPD, QUANTUM) et impliquer les étudiants en médecine pour la saisie et la structuration des données (notamment via les comptes rendus).	Professionnaliser les acteurs et harmoniser les pratiques	Moyen terme	Universités, LEEM, SNITEM, NUMEUM, SIDIV, AFCROs, Sociétés savantes

Observatoire qualité	Mettre en place un observatoire national pour suivi de la qualité, analyse comparative et diffusion des bonnes pratiques.	Suivi continu et diffusion des standards	Moyen terme	Ministère Santé, HDH
Financement pérenne (au-delà du financement initial)	Intégrer la question du financement de la qualité (pour accompagner la structuration des EDSH, le monitoring des registres/cohortes, créer des outils mutualisés...) dans les réflexions déjà engagées sur les modèles de soutien aux producteurs de DSVR	Garantir les ressources pour mise en conformité et monitoring	Moyen terme	Ministère Santé, Industrie, Europe (EEDS), investisseurs privés
Standardisation REQUEST	Généraliser l'outil REQUEST pour registres/cohortes (mise à jour annuelle, intégration au catalogue national).	Standardiser les informations et faciliter l'évaluation	Court terme	HAS, ARIIS, producteurs de données
Monitoring mutualisé	Créer un réseau national d'ARC pour data-management sur site (registres/cohortes), avec tarifs négociés et formation des centres.	Mutualiser les coûts et renforcer la qualité	Moyen terme	Associations académiques, AFCROs

Table des matières

1-	Présentation du CSF IA et santé et du groupe de travail.....	7
	Le CSF.....	7
	Le Groupe de Travail "Exploiter et valoriser les données de santé en vie réelle"	7
	Note méthodologique	9
2-	Problématique	13
	Contexte et enjeux de l'accès à l'information sur la qualité des données	13
	La nécessité d'une approche différenciée selon les types de bases de données.....	13
	La fragmentation des recensements existants.....	14
3-	Une enquête pour objectiver les pratiques	16
4-	Vers une harmonisation européenne : l'Espace Européen des Données de Santé et le projet QUANTUM ¹⁰	18
5-	Solutions envisagées pour relever ces défis.....	20
6-	Conclusion	26
7-	Références	29
8-	Annexes	30
	ANNEXE 1 - GLOSSAIRE.....	30
	ANNEXE 2 - ABSTRACT.....	31
	ANNEXE 3 - MEMBRES DU GROUPE DE TRAVAIL « EXPLOITER ET VALORISER LES DONNEES DE SANTE EN VIE REELLE ».....	32
	ANNEXE 4 - RESULTATS DE L'ENQUETE	33
	ANNEXE 5 - PROJET QUANTUM	38

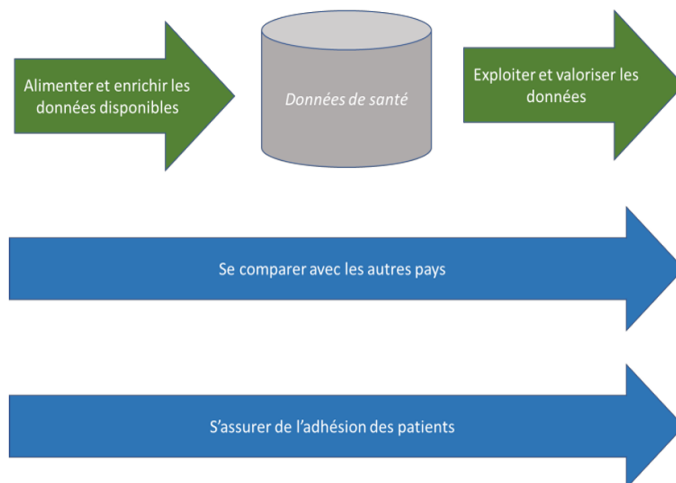
1- Présentation du CSF IA et santé et du groupe de travail

Le CSF

Les comités stratégiques de filière (CSF) ont pour mission de favoriser un dialogue constructif et régulier entre l'État, les entreprises et les représentants des salariés sur tous les sujets clés qui permettent la relance industrielle française et ce, filière par filière. Les dernières années ont permis de confirmer le rôle essentiel des industries et technologies de santé. La filière "Industries et Technologies de Santé" comprend l'ensemble des entreprises françaises de recherche, développement, production et de commercialisation dans le domaine de la santé (entreprises du médicament et des technologies de santé). Les enjeux de ce secteur industriel sont nombreux et complexes mais offrent de grandes opportunités pour améliorer les soins de santé ou innover dans la prise en charge et la prévention.

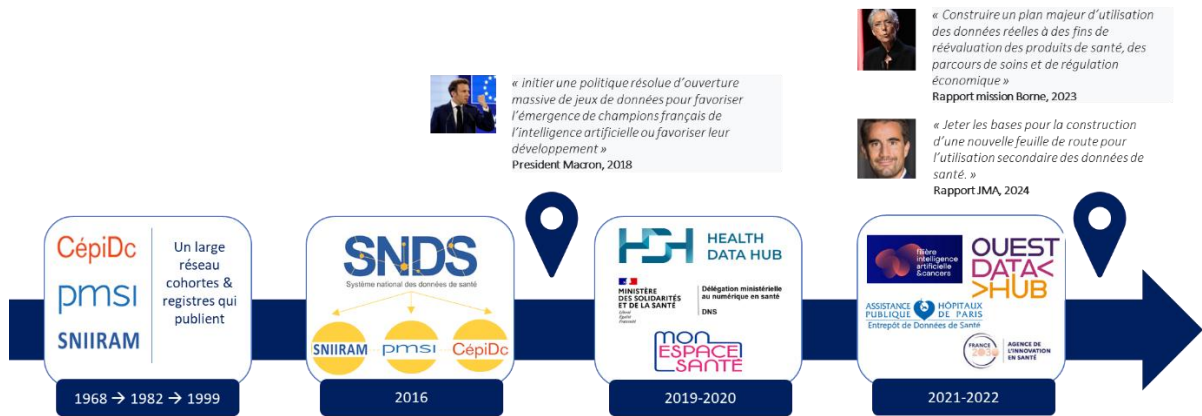
Le contrat stratégique de la filière "Industries et technologies de santé" pour la période 2023-2026 se concentre sur plusieurs enjeux majeurs, parmi lesquels : l'intégration rapide des innovations thérapeutiques et des nouvelles technologies pour améliorer la prise en charge des patients ; le dynamisme de l'innovation et de la recherche à l'échelle internationale ; l'attractivité des compétences et le développement de l'accès aux données et l'optimisation de leur utilisation.

Le projet 12 visant à "Capitaliser sur le patrimoine de données pour favoriser l'attractivité de la France et renforcer sa place dans la compétition internationale" est piloté par Corinne Blachier-Poisson (Présidente, Amgen France) et Mélina Gilberg (Directrice associée - Economie de la santé et génération de données en vie réelle, Amgen France). Il se décline de la manière suivante :



Le Groupe de Travail "Exploiter et valoriser les données de santé en vie réelle"

Parmi l'ensemble des groupes de travail (GT) lancés, l'un d'eux est consacré à l'exploitation et la valorisation des données de santé en vie réelle (DSVR) pour accélérer la recherche. La France est un pays pionnier & dynamique en matière de données de santé, qui s'est fixé des objectifs ambitieux.



Comme pour tout processus d'innovation, la filière se heurte à plusieurs obstacles, dont certains sont recensés ci-après (liste non exhaustive).

	Forces	Marges d'amélioration
Système National des Données de Santé (SNDS)	Quasi-exhaustivité à l'échelle nationale	Enrichir les données médico-administratives avec des données cliniques Accélérer la disponibilité et le rythme de mise à jour des données
Cohortes & registres	Données cliniques structurées	Hétérogénéité dans la qualité des données Hétérogénéité dans les coûts d'accès
Health Data Hub (HDH)	Guichet d'accès unique, considéré comme une référence pour la pré-figuration de l'Espace Européen des données de santé	Simplifier la gouvernance et mieux positionner le HDH dans l'écosystème Enrichir le catalogue sans obliger les bases à « recopier » leurs données sur la plateforme Héberger le SNDS
Mon Espace Santé	98% de la population avec un dossier ouvert ; 300M de documents médicaux téléchargés en 2024	Utilisation par les citoyens (22M de comptes activés à fin septembre 2025), les professionnels de santé (en particulier hors de l'hôpital) Construction de résumés structurés faciliter l'utilisation primaire et envisager éventuellement une utilisation secondaire
Entrepôts de données hospitaliers (EDSH)	Complément de données médicales Mise à jour régulière	Interopérabilité Données non structurées et de qualité très disparate Ressources (humaines, financières) importantes

Ces obstacles constituent des freins majeurs à la réutilisation et la reconnaissance des DSVR.

Pertinence + qualité + simplicité + lisibilité

Fraîcheur	Exactitude	Accès	Evaluation HAS
Représentativité	Validité		Tarifs
Appariement	Exhaustivité		Qualité
	Transparence		Sources disponibles

Ces défis ont déjà été identifiés dans les CSF précédents au travers de recommandations et guides¹, notamment ceux élaborés par le GT « Usage des données de vie réelle dans la prise de décision ».

c. Principales conclusions

Les membres du GT ont identifié six axes de recommandations pour lesquels une série d'actions concrètes pourrait être mise en œuvre :

- **Recommandation n°1 : Renforcer à la source la qualité des données de santé en vie réelle**
 - *Action à mettre en œuvre en priorité* : Soutenir le développement d'outils permettant de garantir et maintenir la qualité des bases de données de référence, à savoir : des indicateurs de qualité permettant de documenter et d'évaluer ces données, un registre déclaratif des bases de données de santé en vie réelle et un outil de requêtage permettant aux acteurs de s'orienter entre les bases.
- **Recommandation n°2 : Renforcer la standardisation des données et systèmes d'information pour accélérer l'interopérabilité et l'appariement entre les bases de données de santé en France**
 - *Action à mettre en œuvre en priorité* : Développer des jeux minimums de données cliniques de qualité et en homogénéiser les formats en fonction des pathologies concernées.
- **Recommandation n°3 : Garantir la qualité et la transparence des méthodes utilisées pour analyser les données de santé en vie réelle**
 - *Action à mettre en œuvre en priorité* : Soutenir le partage et la diffusion des connaissances entre les acteurs sur les méthodes d'analyse des données de santé en vie réelle.
- **Recommandation n°4 : Permettre l'accès aux données de santé en vie réelle dans des temps compatibles avec la prise de décision**
 - *Action à mettre en œuvre en priorité* : Augmenter la fréquence de mise à jour des données disponibles dans le SNDS, en les actualisant au minimum chaque trimestre contre une fois par an actuellement.
- **Recommandation n°5 : Assurer la formation des acteurs de la filière sur les données de santé et le numérique en santé**
 - *Action à mettre en œuvre en priorité* : Soutenir le développement et la visibilité des initiatives existantes (Plan de relance, Académie du Numérique, etc.) visant à favoriser les passerelles entre le monde public et privé et à renforcer l'acculturation à la donnée de santé, en particulier celles dédiées aux professionnels de santé et aux usagers du système de santé.
- **Recommandation n°6 : Améliorer la recevabilité et la prise en compte des données en vie réelle par les évaluateurs des autorités de santé**
 - *Action à mettre en œuvre en priorité* : Poursuivre et intensifier le travail en cours pour renforcer l'utilisation des données en vie réelle lors des évaluations des produits de santé, par exemple en renforçant la lisibilité de l'usage des données en vie réelle dans l'évaluation.

Le GT actuel s'est donc donné pour ambition d'actualiser et de préciser les recommandations du précédent CSF IA et Santé et de promouvoir ces recommandations de manière ciblée.

Note méthodologique

Dans le cadre de ce rapport, plusieurs ressources ont été mobilisées et des actions spécifiques ont été réalisées : analyse bibliographique et documentaire, entretiens semi-directifs et réalisation d'une enquête en ligne.

Une analyse exploratoire, non exhaustive, de la littérature existante, visant à identifier et sélectionner les recommandations que le groupe de travail souhaite promouvoir, et qui a permis de mettre en lumière la question de la qualité des données.

Des publications récentes (rapport du Sénat², Publication des Ateliers de Giens³, Recommandations et Guide du CSF IA et Santé¹, Rapport de la mission Borne⁴, Rapport du LEEM⁵, rapport Marchand-Arvier⁶) ont mis en évidence différents sujets autour de l'exploitation et la valorisation des DSVR pour un usage secondaire. Le GT a souhaité se concentrer sur une recommandation ayant un impact

important et actionnable, pour pouvoir la porter de manière efficiente et réaliste. Le sujet de la qualité des données de santé a émergé rapidement dans les échanges du groupe.

Une liste (non exhaustive) des recommandations retrouvées dans ces différentes publications, et ayant pour thème le sujet de la qualité des données est reprise ci-dessous :

- Optimiser la qualité des données à la source au travers de plateformes, d'outils et d'indicateurs adaptés aux usages et contribuer à une démarche d'amélioration continue ;
- Garantir la qualité et la transparence des méthodes utilisées pour analyser les DSVR ;
- Rendre possible la mise en œuvre d'études de faisabilité préalables à des études conduites par les industriels :
 - Améliorer la recevabilité et la prise en compte des DSVR par les évaluateurs des autorités de santé : Poursuivre et intensifier le travail en cours pour renforcer l'utilisation des DSVR lors des évaluations de produits de santé, par exemple en renforçant la lisibilité de l'usage des DSVR dans l'évaluation.

La qualité des sources de DSVR disponibles en France ne semble donc pas toujours optimale et la reconnaissance par les Autorités de Santé des études menées par les industriels reste incertaine⁷.

Des industriels ont bénéficié d'un accompagnement méthodologique par les Autorités pour concevoir leurs études, et des réévaluations favorables ont parfois été soutenues par des analyses en vie réelle, telles que des comparaisons indirectes avec des cohortes historiques, des données issues d'autorisation temporaire d'utilisation (ATU) ou des cohortes particulièrement robustes.

En revanche, lorsqu'il s'agit d'études post-inscription, les analyses menées à partir de bases de données pourtant recommandées par les demandeurs aboutissent souvent à des résultats non concluants pour la réévaluation des produits de santé. Par ailleurs, certaines études incluses dans les dossiers de remboursement, notamment celles visant à évaluer le fardeau de la maladie, ne sont pas retenues par les Autorités en raison de faiblesses méthodologiques, le plus souvent liées à la qualité des données.

Enfin, les exemples à l'international montrent que la France accuse un retard dans l'exploitation de ces données : le retour sur investissement d'une étude conduite en France s'avère très incertain, compte tenu de l'importance de l'investissement local demandé. A force de réglementation trop stricte, de manque de flexibilité, de délais prolongés, et d'absence de réelle volonté politique, de plus en plus d'industriels préfèrent investir dans d'autres pays plutôt qu'en France pour ce type de projets.

L'amélioration de ces 2 aspects est un préalable indispensable pour maintenir l'attractivité de la France dans ce domaine et construire une véritable filière industrielle autour des DSVR :

- **Qualité des sources de données** : les industriels (et tous les utilisateurs de DSVR) ont besoin d'informations fiables sur la qualité des données présentes dans les bases avant d'initier un projet de recherche. Pour répondre à cet enjeu, plusieurs actions pourraient être envisagées :
 - Afin de sécuriser la pertinence scientifique et la qualité des analyses, les études de faisabilité sont fortement recommandées comme prérequis méthodologique à toute étude menée sur données secondaires.

- Construire un « data dictionary » ou référentiel, définissant les variables minimales qu'une base doit contenir pour être considérée comme « de qualité », en fonction de la pathologie étudiée.
- Publier des indicateurs clés de performance sur la qualité des données, adaptés à la pathologie le cas échéant, en s'inspirant des pratiques américaines où trois indicateurs sont validés et appliqués : *Completeness*, *Exhaustiveness* et *Accuracy*.
- **Recevabilité par les Autorités de Santé** : en parallèle, l'acceptation de l'évidence issue des DSVR, dans un contexte international et d'application du règlement européen sur le joint Health Technology Assessment (HTA) et de l'Espace Européen des Données de Santé (EEDS), devient nécessaire. Les attentes en termes de méthodologie d'étude, de données minimales nécessaires pour répondre à la demande formulée (notamment en cas d'étude post-inscription) doivent être claires et précises et constituer un blanc-seing pour une recevabilité ultérieure de l'étude. Le process pourrait être le suivant :
 - Favoriser les rencontres précoces pour présenter un plan de génération d'évidence en vie réelle, à l'image de ce qui se fait dans les autres pays, ou de ce qui se fait couramment avec la Commission d'Evaluation Economique et de Santé Publique (CEESP) ;
 - Promouvoir les échanges entre les Autorités de Santé et les industriels quant à la pertinence du choix d'une base de données ;
 - Prévoir un calendrier compatible avec la réalisation d'études de faisabilité, des études elles-mêmes, notamment lorsqu'elles sont demandées par les Autorités (ex. études post-inscription). Actuellement, pour une étude post-inscription demandée par les Autorités de Santé, les délais imposés pour produire un synopsis ne sont pas compatibles avec la réalisation d'une étude de faisabilité ([Notice de dépôt d'un protocole d'étude post-inscription portant sur un médicament](#)).

Des auditions préliminaires semi-directives d'acteurs de l'écosystème afin d'actualiser ces recommandations et d'identifier plus finement les causes des difficultés rencontrées et les initiatives déjà mise en place sur le terrain

Sept auditions ont été réalisées auprès d'acteurs institutionnels et de producteurs de données de santé (registres, cohortes, entrepôts de données hospitaliers). Les entretiens ont été réalisés avec les acteurs suivants :

- Le HDH
- France Coag
- CARMEN
- LIFEN
- EDSH du Centre Hospitalier Universitaire (CHU) de Nantes
- EDSH d'ELSAN
- La Haute Autorité de Santé (HAS)

Les sujets abordés ont porté sur la qualité des sources de données : référentiels utilisés, caractère optionnel ou obligatoire, nature déclarative ou vérifiée par un tiers, périodicité, études de faisabilité, transparence et outils nationaux de diffusion. Les contraintes et limites pour optimiser cette qualité ont également été discutées. Par ailleurs, les échanges avec la HAS ont également porté sur la qualité

des méthodes dans le cadre des doctrines (usage des données en vie réelle pour l'évaluation, lisibilité des attentes méthodologiques selon l'usage) ainsi que sur les études post-inscription, notamment le défi des délais imposés.

Une enquête en ligne à destination des producteurs de données, pour comprendre la perception des notions de qualité, les initiatives mises en place (« Bonnes pratiques ») et les freins potentiels.

L'objectif de cette enquête était de réaliser un état des lieux des mesures prises par les producteurs de DSVR pour optimiser la qualité de leurs données.

L'enquête a été menée en mai et juin 2025 auprès de deux types de producteurs de données :

- **Registres et cohortes** (bases de données structurées)
- **Entrepôts de données de santé hospitaliers** (EDSH)

L'enquête était à la fois quantitative et qualitative (10 questions à choix multiples et/ou à réponses ouvertes), et non anonyme. Elle a été disséminée via les réseaux du CSF IA et Santé, des Entreprises du Médicament (LEEM), du Syndicat National de l'Industrie des Technologies Médicales (SNITEM) et de l'Alliance pour la Recherche et l'Innovation des Industries de Santé (ARIIS).

Les thématiques couvertes par l'enquête incluaient :

- Disponibilité et accessibilité des informations sur les données
- Gouvernance et procédures d'accès aux données
- Confidentialité et pratiques de partage d'information
- Adaptations selon le type de partenaire (académique vs. industriel)
- Fréquence et modalités de mise à jour et d'extraction des données
- Outils et méthodes de contrôle qualité
- Pratiques d'études de faisabilité et d'évaluation de projets
- Évaluations de représentativité
- Conformité aux réglementations de protection des données (Règlement Général de Protection des Données - RGPD, Commission Nationale de l'Informatique et des Libertés - CNIL)

17 registres et cohortes (bases de données structurées) ont répondu à l'enquêtes, ainsi que **4 EDSH**.

2- Problématique

Contexte et enjeux de l'accès à l'information sur la qualité des données

La qualité des DSVR est un préalable indispensable à leur utilisation secondaire et leur reconnaissance dans les processus de recherche et de décision en santé. Cependant, en France, la transparence et l'accessibilité des informations concernant cette qualité³.

La mise en place de projets d'études utilisant des données de santé en vie réelle se heurte à des difficultés majeures pour les acteurs de la recherche, qu'il s'agisse des industriels du médicament, du dispositif médical, des bureaux d'études spécialisés ou des chercheurs académiques. Cette opacité sur la qualité des données se manifeste à plusieurs niveaux : absence d'informations standardisées sur les modalités de recueil (systématicité, existence et/ou fréquence des contrôles), difficulté pour réaliser des études de faisabilité préalables, et manque de connaissance des indicateurs de qualité pertinents et de guide pratique pour leur implémentation¹.

Par ailleurs, l'insuffisance des études de faisabilité, ainsi que l'absence d'informations sur la présence effective des patients ciblés et sur la qualité des bases présentes dans les divers recensements (HAS, HDH, France Cohortes, Portail Épidémiologie France), constituent des freins majeurs à une exploitation optimale des DSVR. Ces obstacles conduisent les chercheurs, qu'ils soient issus du secteur public ou privé, à faire des choix mal informés lors de la conception et/ou de la conduite de leurs études, ce qui nuit à la valorisation des producteurs de données français sur la scène internationale, et complique la reconnaissance des DSVR comme preuve valable par les autorités décisionnelles.

Les enjeux liés à la qualité des données sont largement documentés dans la littérature. Les documents de référence de la Food and Drug Administration (FDA) et de l'European Medicines Agency (EMA) soulignent par exemple l'importance de critères tels que la complétude, la cohérence, la représentativité, la temporalité et la comparabilité des données pour garantir leur fiabilité et leur pertinence dans la recherche et la prise de décision en santé⁸. De même, des études montrent que la qualité des données dépend de leur exhaustivité, cohérence et exactitude, des attributs essentiels pour éviter les biais et garantir des conclusions fiables⁹.

La nécessité d'une approche différenciée selon les types de bases de données

Il est essentiel de distinguer deux grandes catégories de sources de données dont les caractéristiques et les enjeux qualité diffèrent substantiellement :

Les registres et cohortes (et entrepôts de données de santé thématiques) collectent des données issues du soin pour permettre dans un second temps de répondre à des objectifs scientifiques précis (usage secondaire), avec des protocoles de collecte et des variables structurées (Case Report Forms (CRF)), souvent assistés par l'Intelligence Artificielle. Ils peuvent utiliser des outils standardisés de contrôle qualité. L'évaluation de la qualité doit porter en priorité sur la complétude du recueil par rapport au protocole initial, l'exhaustivité de l'inclusion des patients éligibles, la durée de suivi et la cohérence des données collectées. La représentativité peut faire l'objet d'une évaluation spécifique.

Les EDSH, en revanche, agrègent des données issues des systèmes d'information hospitaliers couvrant de multiples pathologies et situations cliniques. Ces données n'ont pas été initialement collectées à

des fins de recherche mais pour des besoins de soins et de gestion. Les enjeux qualité y sont différents : non-structuration de nombreuses sources de données, impossibilité d'évaluation de la qualité sur données brutes, hétérogénéité des sources et des opérateurs, absence fréquente de vision longitudinale du parcours patient et faible utilisation d'outils standardisés de qualification. L'évaluation doit alors se concentrer sur la disponibilité des informations (en particulier, au sein des variables non structurées), et la représentativité des populations captées.

La fragmentation des recensements existants

Actuellement, plusieurs initiatives tentent de recenser les bases de données disponibles en France :

Le [Portail Épidémiologie France](#) apparaît comme le recensement le plus exhaustif, permettant d'identifier les bases disponibles et pertinentes pour répondre à une question de recherche spécifique. Toutefois, ce portail ne fournit pas d'informations détaillées sur la qualité des données recensées, limitant ainsi sa capacité à guider les choix méthodologiques des chercheurs.

Le [recensement CNIL des EDS](#) offre une vision des entrepôts de données déclarés auprès de l'autorité de protection des données, garantissant leur conformité réglementaire, mais sans information sur leur contenu et leur qualité.

Le [catalogue du HDH](#), qui répertorie à la fois des EDS et des cohortes, reste limité aux bases qui ont accepté de figurer dans ce catalogue. La couverture reste donc très partielle et l'accès aux métadonnées de qualité variable selon les producteurs de données.

Le [recensement de la HAS](#) liste des bases de données et études en vie réelle susceptibles d'être utilisées pour répondre aux demandes de données complémentaires de la HAS pour l'évaluation des produits et technologies de santé. Les porteurs de données qui souhaitent figurer sur cette liste doivent remplir un dossier. Il ne donne toutefois aucune information sur le contenu des bases ni sur la qualité des données qui les alimentent.

Cette multiplication des recensements, sans harmonisation ni complétude, complexifie la tâche des chercheurs qui doivent naviguer entre plusieurs sources d'information aux formats et niveaux de détail hétérogènes.

Vers un recensement exhaustif par une instance unique ?

Face à cette fragmentation, une première recommandation émergente consisterait à confier à une instance unique, légitime et reconnue par l'ensemble de l'écosystème, la mission de centraliser et d'harmoniser le recensement des bases de données de santé françaises.

Un tel recensement unifié devrait intégrer :

- Une identification exhaustive des bases (registres, cohortes, EDS)
- Une description standardisée des caractéristiques de chaque base en fonction de la pathologie concernée
- Des métadonnées harmonisées facilitant la comparabilité
- Une mise à jour régulière garantissant la pertinence de l'information

Du recensement à la qualification : le défi de l'évaluation de la qualité

Si le recensement permet d'identifier le type de données disponibles et leur pertinence théorique pour répondre à une question de recherche, il demeure insuffisant sans information précise sur la qualité effective de ces données. En France, l'absence de connaissance et d'application des indicateurs standardisés de qualité des bases de données constitue un obstacle majeur.

3- Une enquête pour objectiver les pratiques

Face à ce constat, le CSF IA et Santé, dans le cadre de son groupe de travail GT2 « Qualité des données de santé en vie réelle », a conduit une enquête approfondie auprès des producteurs de données français. Les résultats détaillés de cette enquête sont disponibles en Annexe 4. Ils ont été présentés lors du congrès de l'International Society for Pharmacoeconomics and Outcomes Research (ISPOR) en novembre 2025 et le seront à nouveau lors du Congrès de l'Information Médicale EMOIS en mars 2026. Une synthèse des résultats est reprise ci-après.

Synthèse comparative : registres/cohortes vs. EDSH

Dimension	Registres et cohortes (N=17)	EDSH (N=4)
Nature des données	Structurées, collectées prospectivement pour une pathologie définie, une technologie de santé ou en population générale	Hétérogènes, collectées prospectivement dans le cadre du soin, mixtes (structurées et non structurées)
Évaluation qualité globale	Possible via indicateurs standardisés	Impossible - dépend du projet spécifique
Outils qualité standardisés	53% utilisent	N/A - difficile à évaluer hors projet
Protocole accessible	41% open data, 47% après accord de confidentialité	N/A - pas de protocole au sens classique
Indicateurs qualité détaillés	Majoritairement après contractualisation	Calculés uniquement dans le cadre d'un projet
Études de faisabilité	82% systématiques, souvent avant contractualisation	Systematiques Pour un des répondants, en deux niveaux (macro puis approfondie)
Principal défi qualité	Ressources pour contrôles de cohérence, standardisation entre centres	Structuration des données non structurées, qualification continue dans le cadre de pathologies/projets spécifiques
Dynamique d'amélioration	Amélioration par renforcement des contrôles et multiplication des projets	Amélioration par la multiplication des projets et une structuration progressive

Constats transversaux de l'enquête

Points communs

- **Forte conformité réglementaire** : Tous les producteurs de données démontrent une adhérence stricte aux exigences RGPD et CNIL ;
- **Portails de transparence** : Très large adoption (88% registres/cohortes, 100% EDSH) ;
- **Support scientifique** : Forte disponibilité pour accompagner les projets (76% registres/cohortes, 100% EDSH) ;
- **Études de faisabilité** : Largement pratiquées (82% et 100%), avec des modalités différentes ;
- **Confidentialité comme frein** : Citée fréquemment comme raison de non-partage d'information en amont, mais souvent associée à d'autres facteurs (manque de ressources, crainte de nuire à l'attractivité).

Différences structurelles

- **Maturité des outils qualité** : Fort écart entre registres/cohortes (53% d'outils standardisés) et non-applicable pour les EDSH
- **Nature de l'information disponible** : Protocoles et CRF pour les registres/cohortes vs. types de données et procédures d'accès pour les EDSH
- **Possibilité d'évaluation qualité préalable** : Faisable (même si rarement fait) pour les registres/cohortes, impossible globalement pour les EDSH
- **Modèle économique de la faisabilité** : soit facturée au moment de la demande, soit incluse dans un partenariat

Défis communs identifiés

- **Manque de ressources** : Problématique majeure pour les deux types de structures
 - Personnel dédié au data management insuffisant
 - Difficulté à pérenniser les postes
 - Compétences spécialisées difficiles à recruter et retenir dans les structures publiques
- **Absence de standards partagés** : Besoin exprimé de définitions communes de la qualité et d'indicateurs standardisés
- **Tension entre transparence et confidentialité** : Difficile équilibre entre partage d'information et protection des données
- **Représentativité** : Enjeu important mais peu documenté (seuls 29% des registres/cohortes ont réalisé une étude de représentativité)
- **Sensibilisation des utilisateurs** : Commentaire révélateur d'un registre : "Il faudrait déjà que les demandeurs s'y intéressent... Nous avons souvent l'impression que les utilisateurs ne font pas la différence entre une base de données et une autre"

4- Vers une harmonisation européenne : l'Espace Européen des Données de Santé et le projet QUANTUM¹⁰

Les résultats de cette enquête révèlent une hétérogénéité importante des pratiques, un besoin manifeste de standardisation et de ressources supplémentaires. Dans ce contexte, le cadre européen émergent offre une opportunité d'harmonisation et d'amélioration des pratiques nationales.

Le règlement européen sur l'EEDS, publié officiellement au Journal officiel de l'Union européenne le 5 mars 2025, constitue désormais le cadre réglementaire de référence pour l'utilisation secondaire des données de santé en Europe. Il établit les principes et objectifs pour un usage harmonisé des données de santé à travers l'Union européenne, avec une mise en application progressive à partir de mars 2027. L'article 78 de l'EEDS prévoit que les jeux de données mis à disposition via les organismes d'accès aux données de santé (Health data access bodies - HDABs) peuvent être accompagnés d'un label de qualité et d'utilité. Élément crucial : ce label devient obligatoire pour les jeux de données collectés et traités avec un financement de l'Union européenne ou des États membres, garantissant ainsi un niveau minimal de transparence pour les données issues de financements publics.

Un élément central de l'infrastructure EEDS est la mise en place de catalogues nationaux de métadonnées, proposant une description standardisée des jeux de données pour tous les catalogues de santé européens, améliorant l'interopérabilité et l'accessibilité des données.

L'intégration du label qualité QUANTUM (Quality, Utility and Maturity Measured) dans ce catalogue constitue une avancée majeure : elle permettra aux utilisateurs d'évaluer la qualité et l'utilité d'un jeu de données avant même d'initier une demande d'accès formelle, répondant ainsi au besoin exprimé dans l'enquête.

Le projet [QUANTUM](#), financé par l'Union européenne, vise à développer et mettre en œuvre un mécanisme de labellisation visant à définir des indicateurs standardisés de qualité et d'utilité des jeux de données qui sera idéalement adopté dans le futur EEDS (projet qui se déroule de 2024 à 2026). Le travail de consensus a abouti à l'établissement de dimensions essentielles pour mesurer la qualité et l'utilité des données, avec leurs métriques associées (cf. Annexe 5).

Cette dynamique européenne portée par l'EEDS et QUANTUM offre une opportunité pour la France d'harmoniser ses pratiques nationales avec le cadre européen émergent.

- **Standardisation des indicateurs qualité** : QUANTUM fournit un cadre proche de celui établi par la FDA ou l'EMA. Les dimensions définies offrent une base commune pour les producteurs de données.
- **Transparence sur la qualité** : Il est prévu dans le projet QUANTUM une obligation de labellisation pour les données financées publiquement (article 78 EEDS). Cela signifierait une transparence accrue concernant la qualité des données, que l'enquête a montré insuffisante actuellement (majorité des informations qualité transmises uniquement après contractualisation).

- **Catalogue national harmonisé** : Il permettra un recensement des sources de données, répondant à la problématique de fragmentation actuelle entre Portail Épidémiologie France, recensement HAS, futur portail [Fresh](#), HDH, recensement CNIL, etc.
- **Valorisation internationale** : L'adoption des standards QUANTUM positionnerait les producteurs de données français au niveau des meilleures pratiques européennes, renforçant leur attractivité pour les collaborations internationales.

Toutefois, de nombreux défis restent à relever :

- **Adaptation aux EDSH** : Les spécifications QUANTUM doivent tenir compte de la spécificité des entrepôts hospitaliers, où l'évaluation qualité globale est impossible. Le label devrait indiquer des niveaux de qualité différenciés selon les types de données (structurées vs. non structurées).
- **Ressources nécessaires** : L'enquête a clairement identifié le manque de ressources comme l'un des obstacles majeurs. La mise en conformité avec QUANTUM nécessitera :
 - Du personnel dédié au data management
 - Une communication large sur les nouveaux standards
 - Une adaptation des outils existants pour automatiser la génération des nouveaux indicateurs de qualité
 - Du temps pour remplir les auto-évaluations
- **Culture de transparence** : Passage d'une culture de confidentialité (53% des registres/cohortes citent les préoccupations de confidentialité comme raison de non-partage) à une culture de transparence imposée par l'EEDS.
- **Échéance serrée** : Avec une entrée en application en mars 2029, les producteurs de données français disposent de moins de 4 ans pour se mettre en conformité.
- **Financement** : Question cruciale non résolue - qui financera la mise en conformité ? Les producteurs de données ? Les institutions publiques ? Les utilisateurs de données ?

5- Solutions envisagées pour relever ces défis

Recommandations transversales

✓ Définition des indicateurs essentiels de qualité

Suggestion : Organiser des ateliers réunissant producteurs de données, institutionnels et industriels (ex : Ateliers de Giens) afin d'identifier les indicateurs clés pour les cohortes, registres et EDSH :

- Indicateurs communs à toutes les parties
- Indicateurs d'intérêt pour l'une ou l'autre des parties prenantes
 - Exemples pour les industriels :
 - Indicateurs sur la durée de suivi des patients, la fréquence des mesures par patient
 - Indicateurs de complétude sur des variables clé, c'est à dire des variables qui permettront d'évaluer des produits de santé (efficacité, mortalité...) ; à définir en fonction de chaque pathologie étudiée
 - En cas de complétude insuffisante, indicateur sur la possibilité par le producteur de données de compléter les données manquantes, sur le délai / coût nécessaire, la possibilité (ou non) de compléter les données (ex : de Techs ou attachés de recherche clinique (ARC) externes d'une contract research organization (CRO) ou traitement automatique du langage naturel (natural language processing - NLP).
 - Exemples pour les institutionnels et les industriels :
 - Indicateurs de représentativité d'une cohorte (par ex. grâce à un chaînage avec le SNDS facilité par les accès permanents...)

Cette démarche vise à uniformiser les pratiques en définissant des standards communs et en alignant les méthodologies.

✓ Transformation et structuration des données

Suggestion : Structurer les informations dès leur collecte et faciliter leur intégration dans des systèmes compatibles. Cette transformation doit s'appuyer sur des technologies innovantes tout en définissant un référentiel conforme aux standards internationaux (Fast Healthcare Interoperability Resources - [HL7 FHIR](#), Logical Observation Identifiers Names & Codes - [LOINC](#)) afin d'assurer des échanges fiables et sécurisés entre systèmes hétérogènes :

- Développer l'informatisation à la source (par exemple pour complétion automatique de CRF de cohortes spécifiques) et s'assurer de sa fiabilité
- Améliorer l'interopérabilité entre logiciels (hospitaliers ou autres)
- Accompagner de façon adaptée la montée en charge des outils de transformation des données brutes (NLP, IA)
- Définir un référentiel de variables minimales normalisées, conformes aux standards d'interopérabilité (ex. [HL7 FHIR](#), [LOINC](#)), pour garantir l'interopérabilité et permettre un échange fluide des données entre systèmes différents.

✓ Plateforme nationale d'information

Suggestion : Capitaliser sur les initiatives en cours (type [FReSH](#), GREG IHI) pour créer une plateforme nationale recensant l'ensemble des sources de données (registres, cohortes, EDSH...) avec, pour chaque source :

- Les métadonnées standardisées [HealthDCAT-AP](#) et selon les « [FAIR principles](#) » ([guideline-principes-fair-v2.pdf](#))
- Les labels QUANTUM pour les bases éligibles
- Les informations de contact et procédures d'accès harmonisées
- Un moteur de recherche par pathologie, type de données, critères d'éligibilité

Gouvernance : Entité à déterminer dans le cadre de la mise en place de l'EEDS.

✓ **Charte nationale de transparence**

Suggestion : Élaborer collectivement une charte définissant :

- Les informations minimales à partager en open data
- Les informations partageables après accord de confidentialité simple
- Les informations réservées après contractualisation
- Les délais « standards » de réponse aux demandes
- La convention contrat unique pour l'accès aux données des bases de données existantes

Signataires : Producteurs de données, utilisateurs académiques et industriels, instances publiques.

✓ **Formation et professionnalisation du data management, du monitoring et de la conduite de projets**

Suggestion : Créer un ou plusieurs cursus de formation certifiant national en data management, en monitoring pour données de santé en vie réelle, avec :

- Module sur les standards QUANTUM et [HealthDCAT-AP](#)
- Formation aux outils de contrôle qualité
- Méthodologies de data-management
- Aspects réglementaires (RGPD, CNIL, EEDS)
- Nouvelles méthodologies de collecte de données, notamment par NLP et processus robustes de validation des résultats produits par ces outils
- Gestion de projet appliquée aux registres/cohortes

Portage : Universités en partenariat avec l'ensemble des associations professionnelles représentant les producteurs et les utilisateurs de données (Société Française de Statistique SFDS, Société Française de Pathologie SFP, LEEM, SNITEM, Syndicat de l'Industrie du Diagnostic In Vitro - SIDIV, Association Française des CROs AFCROs, Sociétés savantes etc.).

Impliquer les étudiants (en médecine) dans la saisie des données et même le data-management avec de multiples bénéfices :

- Données bien complétées
- Apprentissage de la pathologie facilité pour ces étudiants
- Sujets de thèses possibles
- Apprentissage pratique standardisé du remplissage des comptes rendus, en lien avec la spécialité, pour améliorer la structuration des données dans les EDSH

Ouvrir ces formations aux investigateurs et à d'autres personnels, soignants (par ex. infirmiers) ou non soignants (data manager, data scientists...).

Renforcer la motivation des centres en valorisant l'implication des équipes par une reconnaissance académique et financière, notamment via les publications issues des registres, cohortes ou EDSH.

✓ **Observatoire national de la qualité des DSVR**

Suggestion : Créer un observatoire national chargé de :

- Suivre l'évolution de la qualité des bases françaises
- Publier un rapport annuel d'analyse comparative des bases (anonymisé)
- Identifier et diffuser les bonnes pratiques
- Accompagner les producteurs de données dans leur montée en maturité
- Créer des réseaux d'investigateurs pour standardiser les pratiques

✓ **Mécanisme de financement pérenne**

Un mécanisme de financement pérenne est indispensable pour permettre une réelle amélioration de la qualité des DSVR. Celle-ci ne pourra être atteinte sans des moyens humains et financiers adaptés, qui doivent devenir une priorité, en particulier pour les EDSH. Les ressources nécessaires concernent notamment le data management, le monitoring ou encore l'usage d'outils avancés de fouille de données.

Suggestion : Intégrer pleinement la question du financement de la qualité dans les réflexions nationales (voire à l'échelle européenne) déjà engagées sur les modèles de soutien aux registres, aux cohortes et aux entrepôts de données. Des contributions provenant d'acteurs divers pourraient être envisagées : financements publics dédiés des ministères Santé et Recherche, contribution européenne (fonds EEDS), utilisateurs industriels en lien avec des projets d'étude, investisseurs privés, etc.

Allocation :

- Soutien aux bases sous-dotées pour mise en conformité de leur qualité d'après les normes QUANTUM
- Financement d'études de représentativité
- Automatisation des indicateurs de faisabilité pour réduire la charge administrative des producteurs sans augmenter le temps humain
- Développement d'outils mutualisés
- Formation des équipes
- Fléchage des financements liés aux points SIGAPS des publications

Cette intégration permettrait de garantir un financement public minimal et durable pour les cohortes situées dans des domaines prioritaires (par ex. sur la base des travaux d'horizon scanning menés par la HAS), et ainsi de revenir à des modèles de financement plus équitables, moins dépendants de partenariats ponctuels avec les acteurs industriels.

Pour les registres et cohortes

✓ **Transparence préalable sur les indicateurs clés**

Suggestion : Publier en open data ou après simple accord de confidentialité (sans contractualisation) un ensemble minimal d'indicateurs qualité :

- Nombre de patients inclus par année d'inclusion croisé avec le nombre de patients suivis pour chaque année suivant l'inclusion
- Durée moyenne de suivi par année d'inclusion

- Taux de perdus de vue global et définition des perdus de vue
- Proportion de patients pour lequel le statut vital est connu
- Variables principales du protocole/CRF avec taux de complétude (si pertinent)

Bénéfice : Permet des décisions éclairées avant engagement contractuel, et augmentation des propositions de projets d'études en vie réelle.

- **Adoption généralisée de l'outil REQUEST**

Suggestion : Systématiser le remplissage de REQUEST pour tous les registres et cohortes, avec mise à jour annuelle et publication des tableaux.

Support nécessaire :

- Simplification de l'outil (retours sur la complexité)
- Accompagnement pour la première complétion avec des exemples
- Intégration dans le futur catalogue [HealthDCAT-AP](#)

L'outil REQUEST pourrait être amélioré et enrichi grâce à une qualification continue, en instaurant des délais pour la complétion et le suivi, ainsi qu'en réalisant des audits qualité réguliers. Les recommandations issues du projet QUANTUM permettront également d'améliorer l'outil.

- ✓ **Réseau de monitoring mutualisé**

Suggestion : Créer un réseau national d'associations d'académiques, d'ARC expérimentés pouvant être mobilisés pour réaliser du data-management sur site, avec des tarifs négociés collectivement.

Bénéfice : Mutualiser les coûts, assurer une qualité la plus standardisée possible, et permettre de former les centres producteurs de données.

Ces mesures appliquées aux registres et cohortes permettront une meilleure définition des projets d'études, grâce à une identification précoce des variables d'intérêt et une collaboration renforcée avec les investigateurs.

[Pour les entrepôts de données de santé hospitaliers](#)

- ✓ **Reconnaissance de la spécificité des EDSH**

Suggestion : Définir un cadre d'évaluation qualité spécifique aux EDSH ("gold standard" de définition de la qualité), distinct de celui des registres et cohortes, reconnaissant l'impossibilité d'évaluation globale, et proposant des métriques spécifiques et adaptées.

Contenu du cadre :

- Indicateurs de qualité par type de données (Programme de Médicalisation des Systèmes d'Information - PMSI, biologie, compte-rendu, imagerie)
- Méthodologie standardisée pour les études de faisabilité
 - Macro-faisabilité pour la sélection de la population : étape systématique, rapide et à coût modéré ;
 - Faisabilité plus approfondie le cas échéant
- Transparence sur les outils de structuration disponibles (NLP, annotation)
- Catalogue des domaines cliniques déjà structurés, en corrélation avec la date de la fin de projet.

✓ **Financement de la structuration progressive et de l'acculturation interne**

Suggestion : Programme national de financement pour structuration progressive des données hospitalières par domaine clinique prioritaire.

Mécanisme :

- Lancer des appels à projets pour développer des initiatives de structuration propres à chaque domaine thérapeutique (ex : NLP, ontologies), selon le format des appels à manifestation d'intérêt du HDH ou selon un format Programme hospitalier de recherche clinique (PHRC)/Programme de recherche médico-économique (PRME)
- Financement de datamarts thématiques
- Partage des outils développés entre tous les EDSH participants

Ce financement permettra d'impliquer systématiquement les cliniciens producteurs de données et de former et d'accompagner les services cliniques, notamment pour la qualité des comptes-rendus.

Ainsi, dans un premier temps, une évaluation des besoins en ressources doit être conduite ; cette étude détaillée des coûts de mise en conformité pour différents types de structures et l'exploration des solutions de financement adaptées sera un pré-requis nécessaire à l'établissement de financements pérennes.

✓ **Standardisation des comptes-rendus par spécialité**

Suggestion : Lancer des initiatives par spécialité (sur le modèle [HACRO-HUGORTHO](#) en orthopédie) pour harmoniser les comptes rendus cliniques lorsque c'est possible, et intégrer un module dédié à cette harmonisation dans la formation universitaire des cliniciens (cf. recommandations transversales).

Gouvernance : Pilotage par les sociétés savantes en lien avec le réseau des EDSH, et les formations universitaires.

✓ **Dynamique d'amélioration continue**

La dynamique d'amélioration continue passe par l'utilisation même de l'EDSH, sur le même modèle que l'IA générative : la qualité d'un EDSH s'améliorera au fur et à mesure de son utilisation. L'EDSH doit être en constante évolution.

Suggestions :

- Qualification continue des données via les projets : chaque projet permet de structurer les données dans son domaine clinique
- Développer des méthodologies de gestion des données manquantes
- Mieux qualifier les demandes des partenaires (liste précise des variables d'intérêt)

✓ **Tarification transparente et équitable**

Suggestion : Établir une grille tarifaire nationale transparent et équitable pour les études sur EDSH, incluant l'ensemble des coûts, notamment la phase de faisabilité. Cette grille préciserait les niveaux de prestation :

- Faisabilité macro (données structurées uniquement) : tarif A
- Faisabilité avec structuration manuelle d'un échantillon : tarif B
- Faisabilité avec structuration automatisée d'un échantillon et contrôle manuel : tarif C

Bénéfice : Garantir une budgétisation claire pour les industriels et éviter l'absence de tarification qui pénalise les EDSH sous-dotés.

La mise en œuvre des recommandations et bonnes pratiques dans l'écosystème des données de santé français repose sur deux piliers essentiels : une communication proactive et une dynamique d'amélioration continue. Il est crucial de sensibiliser rapidement l'ensemble de l'écosystème, notamment via des campagnes d'information, des congrès et des événements scientifiques, afin de promouvoir les standards et favoriser leur adoption. Parallèlement, la mise en place d'un mécanisme de suivi et d'amélioration progressive de la qualité et de la structuration des données garantira une évolution durable et cohérente. En s'inscrivant dans le déploiement de l'EEDS et d'autres outils européens (initiative QUANTUM), ces actions permettront de fédérer les acteurs autour d'une vision commune et d'améliorer la qualité des données secondaires en santé sur le long terme.

6- Conclusion

L'enquête menée auprès de producteurs de données français révèle un écosystème des données de santé en vie réelle à la fois dynamique et hétérogène, engagé dans une démarche qualité mais confronté à des défis structurels majeurs.

Principaux constats

1. La problématique de la transparence sur la qualité des données

La majorité des informations détaillées sur la qualité des données (taux de complétude, durée de suivi, données manquantes, fraîcheur) ne sont accessibles qu'après contractualisation, lorsqu'elles sont disponibles, ce qui n'est pas forcément le cas, empêchant une évaluation éclairée lors des phases préliminaires de conception des études. Ce déficit nuit à l'efficacité de la recherche et à l'attractivité internationale des sources françaises.

2. Les ressources

Le sous-financement chronique de la collecte des données et du data management est identifié comme l'obstacle majeur par les deux types de producteurs. Cette situation compromet la mise en place de processus visant à assurer une qualité suffisante et risque de freiner la mise en conformité avec les futures exigences européennes.

3. Une maturité réglementaire élevée mais une qualité insuffisante

Tous les producteurs démontrent une forte conformité aux exigences RGPD et CNIL, une très grande majorité disposant de portails de transparence. En revanche, l'utilisation d'outils standardisés d'évaluation de la qualité reste très hétérogène et l'absence de contrôles de la qualité des données disponibles complique la comparabilité entre sources.

4. Des études de faisabilité pratiquées mais non standardisées

La quasi-totalité des producteurs réalisent des études de faisabilité, témoignant d'une culture de l'évaluation préalable. Cependant, les méthodologies, délais, et tarifs varient considérablement, rendant difficile pour les porteurs de projets l'anticipation des coûts et des délais de leurs études. Pour répondre à ces enjeux, un guide méthodologique intitulé « Guide sur les études de faisabilité pour des projets de recherche dans le cadre d'utilisations secondaires de données de santé » a récemment été élaboré par un groupe d'industriels du secteur pharmaceutique et de détenteurs de données (producteurs ou gestionnaires de bases de données), sous l'égide du LEEM¹¹. Ce guide vise à harmoniser les pratiques, proposer des repères concrets, faciliter la compréhension mutuelle entre industriels et détenteurs de données, et optimiser la qualité et l'efficacité des études grâce à un parcours structuré et des délais réduits.

Une fenêtre d'opportunité européenne

L'entrée en vigueur progressive du règlement EEDS (2025-2029) et le développement du label QUANTUM offrent une opportunité de transformer structurellement l'écosystème français des DSVR. Le cadre réglementaire européen impose désormais :

- **Une obligation de labellisation qualité** pour les données financées publiquement (article 78)

- **Un catalogue de métadonnées standardisé** ([HealthDCAT-AP](#)) pour améliorer la visibilité et la facilité de recherche des bases de données
- **Des indicateurs de qualité** (12 dimensions QUANTUM) comparables à l'échelle européenne
- **Une évaluation de la maturité** des processus qualité des détenteurs de données

Cette dynamique peut aider à initier les améliorations nécessaires identifiées par l'enquête : transparence accrue, standardisation des pratiques, investissements dans les ressources qualité, et reconnaissance internationale renforcée.

Néanmoins, pour permettre le succès d'une telle évolution, quatre conditions doivent être réunies :

- 1. Volonté politique et financement** : Un investissement public et/ou privé significatif est indispensable pour accompagner les producteurs de données dans leur mise en conformité, particulièrement les structures qui n'ont pas ou très peu de financements publics.
- 2. Adaptation au contexte français** : Les spécifications QUANTUM doivent être déclinées en tenant compte notamment de la distinction fondamentale entre registres/cohortes et EDSH.
- 3. Approche collaborative** : La définition des standards nationaux doit associer l'ensemble des parties prenantes (producteurs de données, utilisateurs académiques et industriels, autorités publiques) pour garantir acceptabilité et applicabilité.
- 4. Accompagnement dans la durée** : Au-delà de la mise en conformité initiale, un dispositif pérenne d'accompagnement, de formation, et d'amélioration continue est nécessaire pour inscrire la démarche qualité dans une dynamique durable.

Limites

Ce rapport s'appuie sur des informations déclaratives recueillies lors d'entretiens avec des acteurs de la filière, incluant des producteurs de données de santé, ainsi que sur les réponses à l'enquête. Un biais de sélection est possible : les données analysées proviennent principalement des acteurs les plus engagés, c'est-à-dire des producteurs volontaires ayant accepté de partager leurs pratiques, ce qui peut offrir une vision partielle de la réalité.

Par ailleurs, la notion de qualité des données est intrinsèquement complexe. Elle varie selon la pathologie étudiée, car les variables essentielles diffèrent d'un domaine à l'autre. De plus, le niveau d'exigence doit être gradué et priorisé en fonction des ressources disponibles. Il est souvent préférable de concentrer les efforts sur des critères clés tels que la **fraîcheur des données**, la **complétude** des variables essentielles et la **durée de suivi**, quitte à compenser ultérieurement certains manques (comme la représentativité ou des variables secondaires) par des études complémentaires.

Perspectives

Ainsi, pour favoriser l'attractivité de la France à l'international et la reconnaissance des DSVR françaises comme preuves valables par les autorités décisionnelles, il est essentiel d'initier dès maintenant un échange autour des normes et bonnes pratiques de qualité, en anticipant les exigences européennes tout en respectant les spécificités de l'écosystème national.

Les résultats de cette enquête constituent une première étape objective d'état des lieux. Ils appellent maintenant une mobilisation collective pour transformer les défis identifiés en opportunités d'amélioration, dans une démarche pragmatique, financée, et inscrite dans le temps long de la transformation des pratiques. Une première initiative de mobilisation collective sera lancée en 2026,

dans le cadre des Ateliers de Giens, où une réflexion impliquant académiques, industriels et institutionnels portera sur la valorisation des données de santé en vie réelle pour un usage secondaire via l'optimisation de leur qualité. Les résultats des travaux sont attendus au premier trimestre 2027.

7- Références

1. Groupe de Travail « Usage des données en vie réelle dans la prise de décision ». *Programme « Intelligence Artificielle et Santé » du Comité Stratégique de Filière Industries et Technologies de Santé ; Recommandations du groupe de travail*. 2021. [Recommandations-du-GT-Usage-des-données-en-vie-réelle-dans-la-prise-de-décision_vf.pdf](#)
2. Commission des Affaires Sociales du Sénat. Données de santé : une réforme encore en cours de chargement - Rapport d'information. Juillet 2023. <https://www.senat.fr/notice-rapport/2022/r22-873-notice.html>
3. de Pourvoirville et al. Données et preuves en vie réelle dans l'évaluation des technologies de santé : dans quels cas sont-elles complémentaires, substitutives, ou les seules sources de données par rapport aux essais cliniques ? *Therapies*. 2023. <https://doi.org/10.1016/j.therap.2022.10.067>
4. Mission régulation des produits de santé. Pour un « New Deal » garantissant un accès égal et durable des patients à tous les produits de santé. Août 2023. https://www.economie.gouv.fr/files/files/directions_services/cge/mission_regulation_des_produits_de_sante.pdf
5. LEEM. Positionnement vis-à-vis du rapport sur l'usage secondaire des données de santé. 23 avril 2024. <https://www.leem.org/sites/default/files/2024-05/2024-PP%20Rapport%20usage%20secondaire%20donn%C3%A9es%20de%20sant%C3%A9%20JMA%20vf.pdf>
6. Jérôme Marchand-Arvier et al. Fédérer les acteurs de l'écosystème pour libérer l'utilisation secondaire des données de santé. Décembre 2023. https://sante.gouv.fr/IMG/pdf/rapport_donnees_de_sante.pdf
7. Brottier L. Comment les données de vie réelle sont-elles appréhendées par la HAS lors des soumissions de dossiers d'oncologie à la CT ? Thèse d'exercice Pharmacie université Paris-Saclay 2024. https://upsaclay.focus.universite-paris-saclay.fr/prim-explore/fulldisplay?docid=33PUP_Alma_UNIMARC21295024470006051&context=L&vid=33UDPS_VU1&search_scope=default_scope&adaptor=Local%20Search%20Engine&tab=default_tab&query=any,contains,lauriane%20brottier&offset=0
8. Klimek, P., Baltic, D., Brunner, M., Degelsegger-Marquez, A., Garhöfer, G., Gouya-Lechner, G., et al. (2022). *Quality criteria for real-world data in pharmaceutical research and health care decision-making: Austrian expert consensus*. *JMIR Medical Informatics*, 10(6), e34204.
9. Cardinal Health. (2025). *Ensuring accuracy in HEOR: The critical role of real-world data completeness*. Cardinal Health Real-World Evidence and Insights. <https://cardinalhealth.com/rwe>
10. QUANTUM Project. (2025). *Home - QUANTUM: The health data quality label*. Récupéré de <https://quantumproject.eu/>
11. LEEM. Guide sur les études de faisabilité pour des projets de recherche dans le cadre d'utilisations secondaires de données de santé. Février 2026. <https://www.leem.org/publication/guide-sur-les-etudes-de-faisabilite-pour-des-projets-de-recherche-dans-le-cadre-d>

8- Annexes

ANNEXE 1 - GLOSSAIRE

AFCROs	Association Française des CROs
AMM	Autorisation de mise sur le marché
ARC	Attaché de recherche clinique
ARIIS	Alliance pour la Recherche et l'Innovation des Industries de Santé
ATU	Autorisation temporaire d'utilisation
CEESP	Commission d'Evaluation Economique et de Santé Publique
CHU	Centre hospitalier universitaire
CNIL	Commission Nationale de l'Informatique et des Libertés
CRF	Case report form
CRO	Contract research organization
CSF	Comité Stratégique de Filière
DSVR	Données de santé en vie réelle
EDSH	Entrepôt de données de santé hospitalier
EEDS	Espace Européen des Données de Santé
EMA	European Medicines Agency
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
GT	Groupe de travail
HAS	Haute Autorité de Santé
HDAB	Health data access body
HDH	Health Data Hub
HTA	Health technology assessment
IA	Intelligence Artificielle
INSERM	Institut national de la santé et de la recherche médicale
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
LEEM	Les Entreprises du Médicament
LOINC	Logical Observation Identifiers Names & Codes
NLP	Natural language processing
PHRC	Programme hospitalier de recherche clinique
PMSI	Programme de Médicalisation des Systèmes d'Information
PRME	Programme de recherche médico-économique
QUANTUM	Quality, Utility and Maturity Measured
RGPD	Règlement général de protection des données
SFDS	Société Française de Statistique
SFP	Société Française de Pathologie
SIDIV	Syndicat de l'Industrie du Diagnostic In Vitro
SNDS	Système National des Données de Santé
SNITEM	Syndicat National de l'Industrie des Technologies Médicales

ANNEXE 2 - ABSTRACT

INTRODUCTION

La qualité des données de santé en vie réelle est essentielle pour la recherche et la décision médicale. En France, plusieurs répertoires existent, mais la transparence et l'accessibilité des informations sur la qualité des données restent limitées. Les chercheurs, académiques ou industriels, rencontrent des difficultés pour identifier les bases disponibles et évaluer leur qualité avant tout engagement. Ce manque de visibilité freine l'utilisation des données françaises, limite les partenariats public-privé et ralentit l'innovation. L'entrée en vigueur du règlement européen sur l'Espace Européen des Données de Santé (EEDS) en mars 2025, imposant une labellisation qualité pour les données financées publiquement, renforce l'urgence de promouvoir les bonnes pratiques.

MÉTHODES

Le Comité Stratégique de Filière Industries et Technologies de Santé IA et Santé a mené une enquête auprès de producteurs de données français : 17 registres/cohortes et 4 entrepôts de données hospitaliers (EDSH) ont répondu. Neuf thématiques ont été explorées : accessibilité, gouvernance, confidentialité, adaptations selon le partenaire, mise à jour, contrôle qualité, faisabilité, représentativité et conformité réglementaire.

RÉSULTATS

Des disparités marquées apparaissent entre registres/cohortes et EDSH. 53 % des registres répondants utilisent des outils standardisés de contrôle qualité, contre une utilisation limitée dans les EDSH. 41 % des registres mettent à disposition leurs protocoles en open data, mais peu disposent d'indicateurs de qualité. Dans les EDSH, les données étant recueillies dans le cadre du soin, l'évaluation globale de leur qualité n'est pas fournie, la qualité étant liée au périmètre du projet. 86 % des producteurs de données déclarent réaliser des études de faisabilité, avec des méthodologies hétérogènes. Seuls 29 % des registres ont évalué la représentativité. Le manque de ressources dédiées au data management est identifié comme un obstacle majeur.

CONCLUSION

L'hétérogénéité des pratiques et l'absence d'indicateur harmonisé révèlent un écart avec les standards internationaux. À l'heure où l'EEDS impose une labellisation qualité, un cadre structuré et des ressources dédiées sont essentiels pour garantir la fiabilité et la valorisation des données en vie réelle.

ANNEXE 3 - MEMBRES DU GROUPE DE TRAVAIL « EXPLOITER ET VALORISER LES DONNEES DE SANTE EN VIE REELLE »

Copilotes	
MSD	Aurélie Sylvain
MSD	Lionel Bensimon
W.L. Gore & Associés	Lucie de Léotoing
Membres	
Roche	Valérie Machuron
Sanofi	Cécile Collignon
Merck	Laura Luciani
argenx	Cécile Blein
bioMérieux	Amparo Sanjuan
GE HealthCare	Henri Souchay
Air Liquide Santé	Sarah Alami
Snitem	Armelle Graciet
Health Data Expertise	Magali Lemaitre

ANNEXE 4 - RESULTATS DE L'ENQUETE

Résultats pour les registres et cohortes (n=17 répondants)

Accessibilité des informations

Protocole de la cohorte :

- 41% disponible en open data (bien que souvent sous forme de description succincte)
- 47% partagé après signature d'un accord de confidentialité (NDA)
- 6% non disponible avant contractualisation
- 6% autre (disponible sur demande individuelle)

CRF et liste de variables :

- 41% disponible en open data (souvent description des types de variables sans détail)
- 53% partagé après signature d'un NDA
- 0% non disponible avant contractualisation
- 6% autre (disponible sur demande une fois accord établi)

Indicateurs de qualité détaillés : Les informations précises sur la qualité (taux de complétude, durée moyenne de suivi, nombre de perdus de vue) ne sont fournies qu'après contractualisation pour la grande majorité des bases. La confidentialité est citée par 53% des répondants comme raison principale de ne pas partager ces informations dès les premières discussions.

Adaptations selon le type de partenaire

Les processus de partage d'information varient selon le type de partenaire pour 18% des bases :

- **Partenaires académiques :** accès généralement plus simple, pas toujours de contrat formel requis
- **Partenaires industriels :** restrictions plus fréquentes (18% rapportent l'absence d'accès aux données brutes, nécessité d'adapter le consentement patient pour transmission de données pseudonymisées, interdiction de comparaisons directes entre produits)

Fréquence de mise à jour des données

- **Annuellement :** 53%
- **À la demande :** 24%
- **Mensuellement :** 6%
- **Autre :** 6% (semestriellement)
- 2 non-répondants

Les mises à jour annuelles concernent principalement les rapports d'activité, statistiques générales de la base, grands indicateurs. Les mises à jour à la demande concernent surtout les extractions spécifiques pour des projets, tableaux de bord personnalisés, ou études de faisabilité.

Outils standardisés de contrôle qualité

53% utilisent des outils standardisés, parmi lesquels :

- Cleanweb et programmes R
- eCRF Ennov
- REQUEST (complété pour la HAS)
- Outils internes de monitoring
- Programmes de validation des données avec recours à des CRO pour monitoring sur place

76% sont prêts à remplir l'outil REQUEST utilisé par la HAS, dont 3 l'ont déjà fait. Quelques réticences exprimées concernent la charge de travail importante et la complexité perçue de l'outil.

Études de faisabilité

82% réalisent systématiquement des études de faisabilité avant engagement dans un projet.

Métriques évaluées (par ordre de fréquence) :

- Estimation de la taille de l'échantillon : 65%
- Disponibilité des variables spécifiques : 59%
- Taux de données manquantes : 47%
- Biais potentiels : 41%

Les délais varient de quelques jours pour des études simples à 2 mois pour des analyses complexes. La plupart des études de faisabilité sont contractualisées, avec des tarifs variant selon la complexité (de gratuit pour les partenaires à plusieurs dizaines de milliers d'euros) ;

Support scientifique

76% offrent une expertise pour conseiller sur la conception du projet, incluant :

- Aide sur le design de l'étude
- Définition des critères d'évaluation
- Sélection des périodes d'étude
- Méthodologie statistique
- Identification des biais potentiels

Transparence et conformité réglementaire

88% disposent d'un portail de transparence pour informer les patients des études réalisées.

Conformité réglementaire : forte adhérence aux exigences RGPD, méthodologies de référence CNIL, pseudonymisation, sécurisation des données (serveurs HDS, ISO).

Représentativité

- **41% n'ont pas réalisé d'étude de représentativité**
- **29% ont réalisé une telle étude**, principalement par :
 - Comparaison avec données PMSI nationales
 - Appariement avec le SNDS
 - Comparaison avec registres FRANCIM
 - Validation par comités scientifiques spécialisés
- **24% ont une étude en cours**, souvent via appariement SNDS
- **6% estiment avoir la quasi-totalité de la population d'intérêt**

Mesures suggérées pour améliorer la qualité

Les répondants ont identifié plusieurs leviers prioritaires :

- **Standardisation et méthodologie** : harmonisation des pratiques, définition de standards partagés
- **Transformation des données** : développement de l'informatisation à la source, amélioration de l'interopérabilité
- **Allocation de ressources** : plus de personnel dédié au data management et au monitoring
- **Qualification continue** : mise en place de dates butoirs pour complétion et monitoring, audits qualité réguliers
- **Monitoring sur site** : visites dans les centres collecteurs, formation des investigateurs
- **Meilleure définition des projets** : identification précoce des variables d'intérêt, collaboration renforcée avec les investigateurs
- **Motivation des centres** : indemnisation suffisante, valorisation académique pour encourager l'implication

Résultats pour les entrepôts de données de santé hospitaliers (n=4 répondeurs)

Le principal enseignement de l'enquête auprès des EDSH est **qu'il n'est pas possible d'évaluer la qualité des données d'un EDSH de façon globale**. Comme l'exprime un répondant : "la qualité des données est à apprécier pour un projet en particulier. Cela nécessite un travail d'évaluation spécifique de chaque projet. Impossible à réaliser sur des données brutes."

Cette impossibilité s'explique par :

- La présence de données majoritairement non structurées (comptes-rendus en texte libre)
- L'hétérogénéité des sources de données
- La variabilité de la qualité selon les services et spécialités en charge de la saisie des données
- La dépendance à la question de recherche posée

Accessibilité des informations

Type de données contenues dans l'entrepôt :

- 0% disponible en open data
- 25% partagé après NDA
- 0% non disponible avant contractualisation
- 75% autre (disponible après premier échange / sur demande)

Informations sur la gouvernance et procédures d'accès : La majorité des EDSH sont en cours de développement de leurs sites internet. Les informations sont généralement transmises sur demande ou lors d'entretiens préalables, mais donc rarement disponibles en ligne.

Différences notables avec les registres/cohortes

Contrairement aux registres et cohortes, les EDSH présentent des caractéristiques distinctes :

Pas d'indicateurs de qualité globaux : Les EDSH ne peuvent fournir de métriques générales de complétude ou de qualité. Ces indicateurs ne peuvent être calculés que dans le cadre d'un projet spécifique, une fois les variables d'intérêt identifiées et structurées, au minimum sur un échantillon de patients.

Qualité dépendante du type de données :

- Données PMSI : structurées, relativement fiables
- Données de biologie : souvent au format LOINC, relativement fiables
- Comptes-rendus : données non structurées nécessitant un traitement pour structurer l'information (NLP, annotation manuelle)
- Imagerie brute et données génétiques : rarement intégrées dans l'EDS ; au minimum, les données issues des comptes rendus sont disponibles, mais leur structuration reste limitée.

Fréquence de mise à jour

- **Mensuellement** : 25%
- **Annuellement** : 17%
- **À la demande** : 42%
- **Autre** : 8% (en cours d'implémentation)

Les mises à jour mensuelles concernent généralement les données structurées (type PMSI) sur lesquelles on peut calculer des indicateurs agrégés, ou bien de simples comptages (nombres de patients, de documents, etc).

Outils standardisés de contrôle qualité

Seul 1 EDSH sur 4 déclare utiliser un outil standardisé pour évaluer et refléter la qualité des données.

Les autres rapportent :

- Utilisation de formulaires internes d'évaluation de la qualité par donnée et par projet
- Absence d'outil adapté aux spécificités des EDS
- Qualification continue au fil des projets

Études de faisabilité

100% réalisent systématiquement des études de faisabilité.

Métriques évaluées :

- Disponibilité des variables spécifiques
- Estimation de la taille de l'échantillon de patients
- Biais potentiels

Méthodologie des études de faisabilité :

Un EDSH partage sa méthodologie :

1. **Faisabilité macro** (rapide, gratuite ou peu coûteuse) :
 - Basée sur les données structurées facilement accessibles (PMSI, codes CIM-10, médicaments)
 - Permet d'estimer les effectifs accessibles
 - Délai de réalisation : quelques semaines
2. **Faisabilité approfondie** (contractualisée, coûteuse) :
 - Implique la structuration de données non structurées ou une annotation manuelle
 - Nécessite un échantillonnage et la vérification de la disponibilité des variables
 - Charge de travail parfois comparable à celle du projet complet
 - Délai de réalisation : quelques semaines à 2 mois
 - Coût : minimum 10 000 euros, selon complexité

Support scientifique

100% des EDSH répondants offrent une expertise pour conseiller sur la conception des projets, incluant :

- Aide à la formulation des questions de recherche
- Ciblage des patients (critères d'inclusion/exclusion)
- Identification des données de vie réelle nécessaires
- Conseil sur le design et les aspects méthodologiques
- Identification des biais et des limites anticipables
- Support biostatistique

Transparence et conformité

100% disposent d'un portail d'information patient (obligatoire dans le cadre du référentiel EDS CNIL).

Conformité réglementaire :

- Autorisation CNIL pour tous les EDS (sur la base du référentiel, ou bien autorisation spécifique)
- Respect du RGPD et des méthodologies de référence
- Pseudonymisation, serveurs agréés HDS
- Certifications (HDS, ISO) pour plusieurs EDS

Mesures suggérées par les répondeurs pour améliorer la qualité

Les EDSH identifient des leviers d'amélioration spécifiques à leur contexte :

1. Cadrage et définition :

- Établir un "gold standard" de définition de la qualité pour les EDS
- Développer des méthodologies de gestion des données manquantes

- Mieux qualifier les demandes des partenaires (liste précise des variables d'intérêt)

2. Transformation et structuration des données :

- Développer l'informatisation à la source (par exemple pour complétion automatique de CRF de cohortes spécifiques)
- Améliorer l'interopérabilité entre logiciels hospitaliers
- Accompagner de façon adaptée la montée en charge des outils de transformation des données brutes (NLP, IA)
- Standardiser les comptes-rendus dans certaines spécialités (exemple : réseau HUGO en orthopédie)

3. Ressources et financement :

- **Critique majeure** : "Avoir de la ressource pour s'en occuper. Cela n'est jamais considéré comme prioritaire"
- Allocation de moyens pour le data management
- Financement pour outils de fouille de données

4. Dynamique d'amélioration continue :

- Commentaire significatif : "Un EDS c'est l'inverse de la pile wonder : plus on s'en sert plus il sert - il faut investir sur l'EDS pour lui donner une vitesse minimale de croisière sinon ça s'écroule"
- Qualification continue des données via les projets
- Chaque projet permet de structurer les données dans son domaine clinique

5. Collaboration avec les producteurs de données :

- Impliquer systématiquement les cliniciens producteurs de données
- Former et accompagner les services cliniques, notamment pour la qualité des comptes-rendus.
- Créer des réseaux d'investigateurs pour standardiser les pratiques

ANNEXE 5 - PROJET QUANTUM

Table 5 - List of DQ&U dimensions (final version)

EHDS category	Dimension (by priority)	Dimension definition
Technical quality	Accuracy	Accuracy refers to the degree to which observations correctly describe what it was designed to measure
Technical quality	Precision	Precision refers to the degree of approximation by which data can represent reality
Technical quality	Coherence	Coherence is defined as the dimension that expresses how different parts of the dataset are uniform in their representation and meaning over time, such as formats, semantics (stability of the data models), and methods
Technical quality	Consistency	Consistency refers to the degree to which data has attributes that are plausible and are uniform with other data and over time
Data documentation	Metadata scope	Metadata scope refers to the availability, comprehensiveness, level of detail of metadata and data dictionary that help users understand the data being used
Technical quality	Validity	Validity refers to the degree to which representations of data in a dataset conform to the specification of a data model or data models
Access and provision	Accessibility	Accessibility refers to the dataset being accompanied by clear and transparent access and usage conditions
Data documentation	Compliance	Compliance refers to the degree to which data has attributes that adhere to ethical standards, conventions, protocols or regulations
Coverage	Population representativity	Population representativity refers to the degree to which the data adequately represent the population in question
Coverage	Population coverage	Population coverage refers to the degree to which a dataset includes the potential eligible population
Technical quality	Completeness	Completeness refers to the degree to which all information that could be available is present in a particular dataset
Data documentation	Data provenance	Data provenance means a description of the source of the data, including context, purpose, method and technology of data generation, documenting agents involved in the provenance of data, data validation routines, source data verification, traceability of changes, and quality control of data