Guide sur la qualité des Bases de Données à visée de recherche en santé

Conseils pour qu'une base de données soit exploitable et partageable

Table des matières

1. Introduction

Ce guide est le fruit d'une collaboration active et enthousiaste de plusieurs composantes de F-CRIN dans le cadre du groupe de travail F-CRIN « Qualité des bases de données ».

Ce guide est une des premières étapes de la stratégie scientifique de F-CRIN pour 2020-2024 vers la stratification de patients et pour des essais de médecine de précision. Ce document est amené à évoluer, en particulier avec les recommandations ou exigences qui seront formulées après les premières expériences du Health Data Hub[1].

Les recommandations de ce guide sont à destination des composantes labélisées F-CRIN et elles sont applicables aux bases de données à finalité de recherche en santé. **Les recommandations et conseils ont été pensés pour aider les équipes investigatrices (cliniciens et personnel de leur équipe que ce soit au niveau des coordinations ou des centres d'investigation et de recrutement)** et formulés de façon la plus pragmatique possible par des personnes spécialistes dans leur domaine dans un esprit de "retour d'expérience" qui permet d'éviter des pièges et de tendre vers des bonnes pratiques.

Les bases de données s'entendent ici comme des bases collectant des données individuelles de participants (file active clinique ou cohortes de patients, par exemple) collectées dans le cadre d'un projet de recherche clinique ou du soin. Ces bases regroupent des **données individuelles de participants** centrées sur une pathologie ou une question précise de santé et elles sont recueillies **à visée de recherche** même si elles sont collectées dans le cadre du soin (par exemple : variables cliniques, démographiques, traitement, biologie, imagerie, etc…) par une des composantes de l'infrastructure F-CRIN.

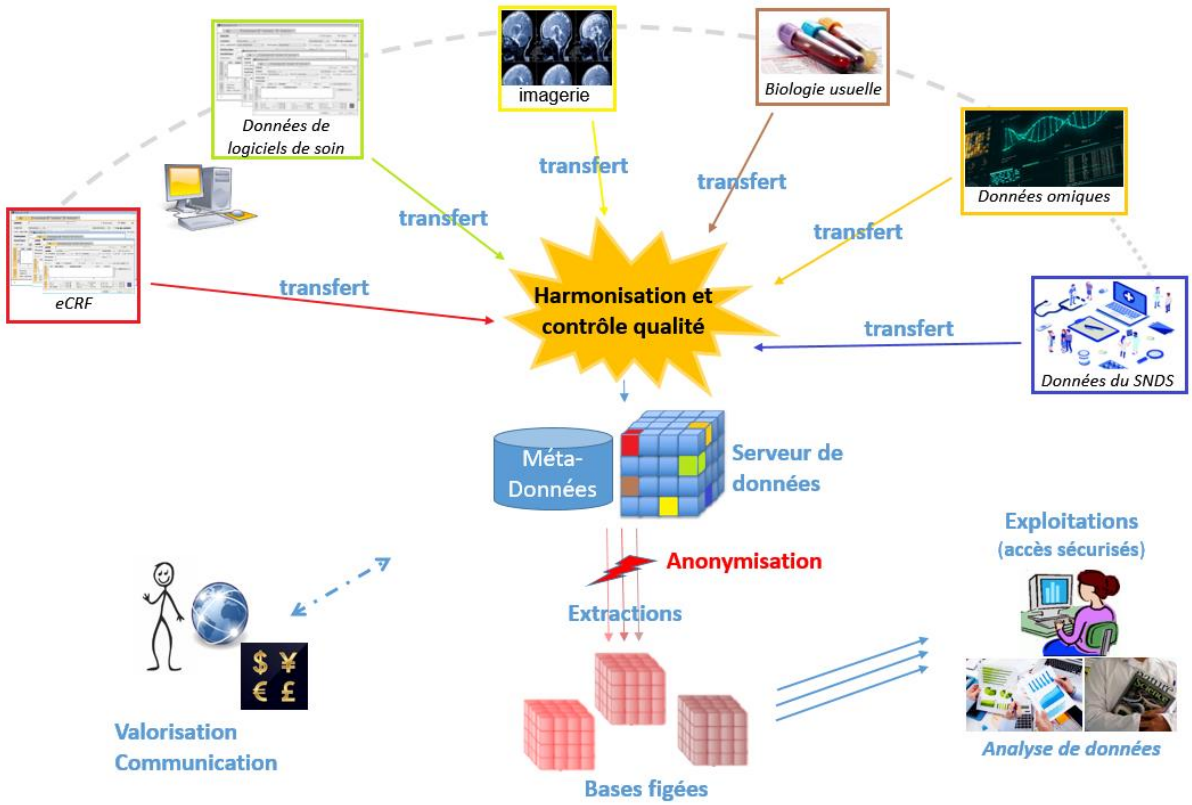

Figure 1 : Représentation d'une base de données multimodale : interopérabilité des sources de données

[1] Plateforme d'exploitation des données de santé françaises, créée par la loi relative à l'organisation et à la transformation du système de santé, à l'initiative Agnès Buzyn, ministre des Solidarités et de la Santé.

Ces données pouvant être de sources différentes *(voir Figure 1)* sont collectées dans une perspective de **partage et/ou de ré-utilisation secondaire**. Ainsi, ces bases de données favoriseront la stratification de patients et permettront d'améliorer la qualité et l'efficience des essais cliniques.

Le but n'est pas seulement d'avoir des "Big data" (données massives en termes de volume, vélocité et variété)[2] mais des "FAIR data" (Findable, Accessible, Interoperable, Reusable data)[3] pour la recherche en santé, particulièrement maintenant que le « Health Data Hub » est lancé.

Ce document définit l'ensemble des spécifications opérationnelles des différents composantes d'une base de données de qualité. Il fournit des lignes de bonne conduite essentielles, applicables dès la conception du projet d'une base de données à vocation de recherche, mais aussi lors de la collecte des données et jusqu'à l'analyse de ces données. Ce guide opérationnel/pragmatique fixe les grands principes pour obtenir des bases de données de qualité mais il ne remplace pas les guides et les recommandations plus détaillés auxquels doivent se soumettre les professionnels du data management et de la recherche clinique.

➔ *Pour toutes suggestions d'amélioration de ce guide, veuillez les transmettre à l'adresse contact@fcrin.org.*

2. Principales définitions et abréviations

Anonymisation	Désigne le processus permettant de rompre tout lien entre une donnée et une personne, afin de produire une donnée anonyme. Le plus souvent en recherche clinique, nous analysons des données pseudonymisées et non des données anonymes.
Analyse d'Impact relative à la Protection des Données	Document d'évaluation permettant d'apporter la preuve du respect des principes fondamentaux de la protection des données personnelles et la bonne gestion des risques pour les participants liés à la sécurité des données (preuve de conformité au RGPD). Cette analyse d'impact est obligatoire pour les traitements susceptibles d'engendrer des risques élevés pour les droits et libertés des participants. Elle est requise pour tous les traitements de données sensibles ou de données hautement personnelles (Données directement identifiantes, NIR (Numéro d'Inscription au Répertoire, appelé couramment numéro de Sécurité sociale), dès lors qu'elles portent sur des personnes vulnérables (patients, personnes âgées, enfants…))[4]. La CNIL requiert systématiquement une analyse d'impact pour les traitements portant sur des **données génétiques** relatives à des patients. L'AIPD (nommée également DPIA : Data Protection Impact Assessment, ou PIA : Privacy Impact Assessment) se décompose en trois parties : • Une description détaillée du traitement mis en œuvre, comprenant tant les aspects techniques qu'opérationnels,

[2] En référence aux « 3 V » du Big Data : https://www.journaldunet.com/solutions/expert/51696/les-3-v-du-big-data---volume--vitesse-et-variete.shtml

[3] Facile à trouver, Accessible, Interopérable et Réutilisable

[4] Pour savoir comment procéder pour savoir si votre traitement requiert une analyse d'impact : https://intranet.inserm.fr/securite-et-prevention/protection-donnees-personnelles/reglementation-generale/Pages/formalites.aspx
ou https://www.cnil.fr/fr/ce-quil-faut-savoir-sur-lanalyse-dimpact-relative-la-protection-des-donnees-aipd

	• L'évaluation, de nature plus juridique, de la nécessité et de la proportionnalité concernant les principes et droits fondamentaux (finalité, données et durées de conservation, information et droits des personnes, etc…) non négociables, qui sont fixés par la loi et doivent être respectés, quels que soient les risques, • L'étude, de nature plus technique, des risques sur la sécurité des données (confidentialité, intégrité et disponibilité) ainsi que leurs impacts potentiels sur la vie privée, qui permet de déterminer les mesures techniques et organisationnelles nécessaires pour protéger les données.
Base de données	Ensemble d'informations structurées mémorisées sur un support permanent et accessible par des moyens électroniques ou d'une autre manière. Une base de données s'appuie sur un Système de Gestion de Base de Données (SGBD).
CDASH	Clinical Data Acquisition Standards Harmonization
CNIL	Commission Nationale de l'Informatique et des Libertés. Créée en 1978, la CNIL est une autorité administrative indépendante qui exerce ses missions conformément à la loi Informatique et Libertés du 6 janvier 1978 modifiée par la loi du 20 juin 2018. Dans l'univers numérique, la Commission Nationale de l'Informatique et des Libertés (CNIL) est le régulateur des données personnelles. Elle accompagne les professionnels dans leur mise en conformité et aide les particuliers à maîtriser leurs données personnelles et exercer leurs droits.
CRF / eCRF	Case Report Form / electronic Case Report Form (anglicisme de « Cahier d'Observation électronique »)
Délégué à la protection des données (DPD) / **Data Protection Officer** (DPO)	Pilote la démarche du responsable des traitements de données personnelles en vue de se conformer aux règles sur la protection des données (règlement européen, droit national, règles internes). Désignation obligatoire pour les organismes publics, les organismes dont l'activité principale est d'effectuer des traitements à grande échelle de données personnelles de santé ou qui suppose un suivi à intervalles périodiques des personnes concernées. Le DPO compétent pour délivrer des conseils sur le montage ou l'utilisation d'une base de données est par définition le DPO de l'institution responsable de cette base (exemple, l'INSERM ou un CHU en particulier) et le Référent à la Protection des Données (RPD) qui assure une mission d'appui de proximité au sein des structures.
Donnée	Information brute ou interprétée concernant un individu (ou plusieurs individus : données agrégées).
Donnée à caractère personnel	Toute information identifiant directement ou indirectement une personne physique (ex. nom, numéro d'immatriculation, numéro de téléphone, photographie, date de naissance, commune de résidence, empreinte digitale, etc...).
Donnée anonyme	Donnée qu'il est impossible de rattacher à une personne donnée. Elles ne sont pas soumises au RGPD, ni à la loi "Informatique et Libertés" qu'elles soient anonymes initialement ou après une anonymisation par un processus permettant de garantir que la personne concernée ne pourra pas être ré-identifiée par la suite (données anonymisées). Travailler sur des données anonymes ou anonymisées permet donc de s'affranchir des dispositions du RGPD et de la loi Informatique et Libertés.

Données pseudonymisées	Les données sont dites pseudonymisées lorsque celles-ci ne peuvent plus être attribuées à une personne précise sans avoir recours à des informations supplémentaires (clés de ré-identification). Ces « clés de ré-identification » doivent être « conservées séparément des jeux de données stockées dans la base de données et soumises à des mesures techniques et organisationnelles » permettant de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique (Ex : recours à une table de correspondance entre le jeu de données pseudonymes (codées) nécessaire aux analyses et les données d'identité conservées séparément dans les centres investigateurs, classiquement utilisée dans les essais cliniques). Dans la mesure où les données pseudonymisées restent rattachées à la personne concernée par un identifiant, ce sont des données à caractère personnel soumises à l'application de la réglementation relative aux données personnelles, à la différence des données anonymisées pour lesquelles les clés de ré-identification ne sont plus disponibles. D'autre part, une donnée pseudonymisée est non nominative mais elle peut rester identifiante (exemple : une date de naissance complète, couplée à d'autres variables telles que le lieu de naissance, la commune de résidence actuelle ou des caractéristiques cliniques, etc…).
Données sensibles	Ce sont des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique. Ces données font l'objet d'une protection renforcée.
DQM	Data Quality Management, anglicisme de Gestion de la Qualité des Données
DTA	Data Transfer Agreement, anglicisme pour Contrat de Transfert des Données
FAIR-data	Findable, Accessible, Interoperable and Reusable data, anglicisme de : Facile à trouver, Accessible, Interopérable et Réutilisable
Figer une base	Gel intermédiaire de la base (différent du gel de base qui est fait après la saisie de l'ensemble des données attendues pour un projet), réalisé de préférence après que les contrôles qualités soient effectués.
Masque de saisie	Espaces dans l'interface utilisateur du logiciel de saisie pouvant comporter plusieurs zones : cases pour saisie de chiffres ou de texte, boutons radio, combo-box, cases à cocher, listes, boutons. Le masque de saisie permet une interactivité entre l'utilisateur et la base de données, homogénéise le format des entrées et en évalue éventuellement la cohérence.
Outil de recueil	Support permettant la collecte des données : par exemple CRF papier, eCRF.
PAQL	Plan d'Assurance Qualité du Logiciel
Participant	Personne volontaire saine ou malade ayant accepté de participer à la cohorte, au registre, à la recherche clinique ou au recueil de données, et dont les données sont collectées afin d'alimenter la base de données.

Responsable de traitement	Personne physique ou morale, autorité publique, service ou organisme qui « détermine les finalités et les moyens du traitement » sauf désignation par une disposition législative ou réglementaire. C'est sur lui que pèse la responsabilité juridique du traitement des données. Le responsable de traitement est également le promoteur lorsque la recherche implique la personne humaine.
RGPD	Règlement Général sur la Protection des Données, il encadre le traitement des données personnelles sur le territoire de l'Union européenne et au-delà dès lors que l'on cible des résidents européens depuis le 25 mai 2018. Ce règlement européen s'inscrit dans la continuité de la Loi française Informatique et Libertés de 1978 et renforce le contrôle par les citoyens de l'utilisation qui peut être faite des données les concernant.
RIPH	Recherches Impliquant la Personne Humaine
Thésaurus	Liste organisée de termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance.
Traitement de données à caractère personnel	Toute opération, ou ensemble d'opérations, portant sur des données à caractère personnel, quel que soit le procédé utilisé (automatisé ou non) : collecte, enregistrement, organisation, conservation, adaptation, modification, extraction, consultation, utilisation, communication par transmission diffusion ou toute autre forme de mise à disposition, rapprochement ou interconnexion, verrouillage, effacement ou destruction, etc...
Variable	Caractéristique (âge, salaire, sexe, glycémie...) définie sur la population et observée sur l'échantillon. Si la variable est un nombre réel, elle est dite quantitative (âge, salaire, taille...), sinon elle est dite catégorielle ou qualitative (sexe, catégorie socioprofessionnelle...). Si les modalités d'une variable qualitative sont ordonnées (i.e. tranches d'âge), elle est dite qualitative ordinale et sinon qualitative nominale.

3. Les 10 conseils essentiels

1- Bien anticiper les aspects de gestion de données en amont, quitte à retarder un peu le début du projet,
2- Associer au projet un personnel de recherche clinique, type data manager, le plus tôt possible,
3- Bien identifier le logiciel informatique et la solution d'hébergement des données en fonction des conditions d'acquisition des données, de stockage ; utiliser des moyens de transfert de données sécurisées pour le partage,
4- Réfléchir en amont aux besoins de réutilisation des données, de partage, de transfert, etc… et prévoir un consentement et une notice d'information qui permettent cela,
5- Etablir un contrat cadre entre les partenaires dès le commencement (par exemple accord de consortium avec chapitre spécifique sur la gestion des données),
6- Respecter les aspects réglementaires et juridiques (RGPD, CNIL, autorisations éthiques, etc… ; nécessité d'un promoteur légal ; aspects de propriété et confidentialité des données qui doivent en général être pseudonymisées),
7- Utiliser tout au long du projet et dans les différentes composantes (CRF, biobanque, questionnaires, etc…) un même identifiant unique pour une personne donnée (dans la mesure du possible),
8- Standardiser le format des variables recueillies,
9- Assurer la véracité et le contrôle qualité des données (rédiger un plan de data management et un plan de monitorage),
10- Assurer la traçabilité et la reproductibilité (toujours dater, tracer, signer toute intervention ou modification sur les données).

Nous vous recommandons de porter une attention particulière aux items suivants :

- Consentement et Notice d'information : « partage et réutilisation des données » *(page 7)*,
- Architecture et support électronique (logiciel) *(page 9)*,
- Identification des participants *(pages 9 et 10)*,
- Variables *(page 10)*,
- Format des variables *(page 11 et 12)*,
- Figer/Geler une base de données pour une analyse *(page 13)*.

4. Critères de qualité

a. Prérequis

Respect Réglementation Recherche Clinique	Le contexte réglementaire en vigueur doit être anticipé et respecté. En effet, ces « regroupements de données de participants » sont encadrés par la réglementation relative à la recherche clinique : **Loi Jardé, applicable aux RIPH, à articuler avec le RGPD et la loi « Informatique et libertés », et la loi « Touraine »** qui créé le Système National des Données de Santé (SNDS) et en organise l'accès avec des exigences de sécurité accrues. Ces aspects de la réglementation seront, bien entendu, gérés par le promoteur (s'il y en a un) de la base de données de recherche, responsable du traitement. **Les Bonnes Pratiques Cliniques sont évidemment applicables.**
Consentement et Notice d'information : « partage et réutilisation des données »	➔ *Pour plus d'informations, se reporter au chapitre **"7.a.2 Consentement et Notice d'Information"**.* Dans une logique de transparence, il est indispensable d'**anticiper le devenir des données personnelles** collectées en mentionnant dans la note d'information et le formulaire de recueil du consentement des participants (s'il est utilisé, ce qui n'est pas le cas des recherches avec « non-opposition du patient »), le partage futur des données et leur réutilisation (y compris à l'étranger, dans des pays respectant un niveau adéquat de protection des données, *cf. chapitre 7.b. Transfert de données*) pour **une ou plusieurs finalités de recherche** en santé (ces dernières peuvent être plus ou moins larges mais suffisamment explicites pour en assurer une bonne compréhension par la personne concernée). *NB : L'information générale ne dispensera pas de l'information individuelle préalable spécifique à chaque nouveau projet de recherche qui nécessite la réutilisation secondaire de données déjà collectées.* Penser, dans la note d'information initiale, à renvoyer vers un **dispositif spécifique d'information,** tel qu'un site internet, auquel les personnes pourront se reporter, pour ne pas avoir à contacter directement et personnellement la personne avant la mise en œuvre d'un nouveau traitement, conformément à la méthodologie de référence MR004. *Si des données sont déjà collectées dans les bases que l'on souhaite partager ou réutiliser et que cette possibilité n'a pas été anticipée dans l'information initiale, les participants concernés devront recevoir une nouvelle information et être en mesure de s'y opposer.* [5]

[5] Faute d'avoir anticipé, il sera possible de demander une dérogation à l'obligation d'information mais :

- L'absence d'information des personnes rendra la recherche non éligible aux méthodologies de référence de la CNIL.

Protocole	Comme ces bases de données ont une finalité de recherche, il est nécessaire **d'écrire au préalable un protocole,** même simple et concis.
Documents-source	Toutes les données entrées dans ces bases doivent être issues de documents-source. En effet, en l'absence de document-source, il n'y a pas de correction de saisie erronée possible. *Cas particulier : il existe des situations où la saisie de la donnée est directe, sans document-source : par exemple les questionnaires en ligne ou sur tablette, les données « images », etc…*
Traçabilité / Reproductibilité	À chaque étape de la « vie » de la base de données il faut penser à la traçabilité et la reproductibilité (définition des variables, évolution de la base, extraction, transfert, etc…). Les informations suivantes devront être documentées et associées à la base de données pour en permettre l'exploitation par d'autres personnes que celles qui l'ont élaborée : • Un **catalogue de variables** avec leur nom, leur intitulé, leur format, l'unité de mesure, la méthode de dosage si besoin, les thesaurus utilisés, • Le **PAQL** (Plan d'Assurance Qualité du Logiciel) du programme qui a permis de produire la base pour analyse (code commenté, architecture technique, etc…), • Le **DQM** (Data Quality Management) pour lister les modifications apportées aux données brutes en vue de l'analyse *(cf. chapitre 4.f. Analyse Statistique)* • **Plan de gestion de données** / Data Management Plan. • **Plan de monitorage** : un document spécifique décrira les procédures mises en œuvre pour vérifier que les données recueillies sont en accord avec la réalité et correspondent à des données de **participants "réels"** (et non complètement inventées) ; à minima, un **plan de monitorage (ou d'autre procédures)** permettant d'attester qu'un consentement a été signé et que l'existence du participant a été vérifiée. Les procédures mises en œuvre pour ces vérifications peuvent dépendre du risque de la recherche et de ses retombées potentielles et utiliseront volontiers des innovations techniques pour des vérifications à distance.

b. Aspects déterminants dès la conception du projet

Hébergement et sauvegarde	Utiliser un **serveur centralisé, bien compartimenté,** et en cas de recours à un **prestataire d'hébergement**, veiller à ne faire appel qu'à des organismes présentant des garanties suffisantes : • Evaluer le degré de maturité du sous-traitant, • Rédiger des cahiers des charges adéquats,

- Le responsable du traitement devra donc effectuer une demande de dérogation à l'obligation d'information dans le dossier de demande d'autorisation qui devra être adressé à la CNIL.

Cette demande, pour être justifiée, devra être accompagnée d'un argumentaire démontrant :

- Qu'il est effectivement impossible de communiquer les informations à la personne concernée ou que l'information exigerait des efforts disproportionnés ou que le simple fait d'informer les personnes compromettrait les objectifs de la recherche.
- Toutes les actions engagées pour informer les personnes qui peuvent l'être, dans la mesure où elles peuvent l'être.

Faute d'être en mesure d'effectuer une information individuelle, le responsable du traitement devra rendre l'information accessible au public, par exemple en mettant les informations sur son site internet ou en les publiant par voie d'affichage.

- **Formaliser la prestation, préalablement, dans un contrat** définissant les caractéristiques du traitement, ainsi que les différentes obligations des parties en matière de protection des données (article 28 du RGPD).

L'agrément ou la certification comme **« Hébergeur de Données de Santé »** de l'hébergeur est la garantie d'un haut niveau de sécurité[6]. Toutefois, en règle générale pour les bases de données constituées à des fins de recherche contenant des données pseudonymisées, il n'est pas requis par les textes que l'hébergeur soit agréé « hébergeur de données de santé » sauf si les données servent aussi à des finalités de soins.

Bien définir en amont le **propriétaire de la base** (qui sera celui qui bénéfice de la *« propriété intellectuelle »* de la base, à différencier de l'hébergeur qui peut être un sous-traitant par exemple), le mode de **sauvegarde** et les **accès** autorisés et ses modalités.

Bien définir aussi les **transferts/exports de données** du serveur hébergeur vers le serveur d'exploitation applicable.

Architecture et support électronique (logiciel)

L'architecture de la base et le support électronique qui seront utilisés doivent être **définis à l'avance.** Dans ce but, il faut impliquer au plus tôt dans le projet, un **personnel de recherche clinique** qui connaît la gestion des données.

L'architecture de la base et le choix du logiciel sont en partie liés, reposant sur un **recueil de données structuré** : plusieurs tables, reliées entre elles, organisation et hiérarchisation des variables… Il est préférable de choisir un logiciel intégrant directement des outils de data management.

Exemple de logiciel de base de données : SAS, ACCESS, SQL server, RedCap, Ennov Clinical, etc…

Evolutivité de la base

Il est préférable de construire une **base modulaire et adaptative**, pour pouvoir, si besoin, ajouter des recueils supplémentaires, la fusionner avec d'autres bases, etc…

Faire évoluer les bases uniquement quand cela est vraiment nécessaire et de façon périodique (c'est-à-dire grouper les modifications nécessaires) est préférable à des modifications au fil de l'eau (ces dernières sont sources d'erreurs, chronophages, peuvent engendrer des complications pour l'analyse, mais aussi pour la gestion des données).

Identification des participants

Les bases utilisées pour l'analyse ne doivent pas contenir de **données nominatives** autant que faire se peut (il n'en reste pas moins que certaines variables ou la conjonction de certaines variables peuvent être identifiantes).

Il faut motiver systématiquement la pertinence de la collecte **de données sensibles et de données d'identification directe** (noms, prénoms, adresse postale ou électronique nominative) qui donnera lieu à un **examen particulier de la CNIL** (lors de l'instruction du dossier ou en cas d'inspection). D'autre part, dans ces conditions la recherche ne sera pas éligible à l'application d'une méthodologie de référence.

Il est primordial de décider d'un **identifiant unique par participant**, de conserver une **liste de correspondance** et de **documenter l'identifiant du participant.**

*La pseudonymisation des données est un bon compromis entre les besoins de la recherche et la sauvegarde des intérêts des participants en **limitant la gravité des impacts potentiels**, notamment en cas d'atteinte à la*

[6] Liste des hébergeurs agréés disponible à l'adresse suivante : https://esante.gouv.fr/labels-certifications/hds/liste-des-herbergeurs-agrees

	*confidentialité des données. **L'impossibilité** de pseudonymiser les données doit rester **exceptionnelle** et elle doit être **documentée**.* *Exemple simple : centre + numéro participant (ordre séquentiel) +/- initiales du participant (qui correspondent à la première lettre du nom et la première lettre du prénom)* *Attention : les initiales du participant contribuent à un potentiel d'identification mais elles peuvent être supprimées avant un transfert des données à l'extérieur ou remplacées par un code lettre qui ne correspond pas aux initiales du participant.*
Variables	Chaque variable doit être "**utile**" et contribuer à l'analyse ou correspondre à une exigence réglementaire. *Pour exemple, les variables biologiques de "routine" ne méritent pas toutes d'être enregistrées.* *Certaines variables servent plus au suivi du recueil ou de l'étude qu'à l'analyse. (Cf. chapitre "4.d. Les variables")*
Financement	Des ressources financières sont nécessaires pour : la création, l'alimentation et la maintenance de la base de données et du serveur. Une activité de **Data management** est nécessaire et plus particulièrement encore pour des bases multi-modulaires (plusieurs hiérarchies de données, différents types de données) ou pour des bases avec des enjeux de pharmaco-Epidémiologie, d'aide à la décision médicale ou de Santé Publique. Pour ces bases spécifiques (et idéalement pour tout type de base), il faut donc intégrer du temps et du financement RH pour ce type de métier.

c. Outil de recueil/saisie

Contributeurs indispensables à la conception	La conception du masque de saisie, le choix et le format des variables doivent être **co-construits** avec un **Data manager/Informaticien** et un **spécialiste de la pathologie** concernée, ainsi qu'avec un **spécialiste de leur analyse ultérieure** (statisticien). *Penser l'outil de saisie en fonction du personnel qui va assurer la saisie (le personnel le moins expert impose le rythme).* Favoriser une collecte/saisie des données par un **personnel de recherche clinique formé.**
Cahier des charges de l'outil	Favoriser un outil avec des **masques de saisie structurés** (un logiciel tableur (de type Excel) ne rentre pas dans cette définition). Il faut s'orienter vers un outil efficace permettant de paramétrer aisément chaque variable (si possible en open source car il est gratuit et le code informatique est publiquement disponible et modifiable par la personne qui se l'approprie) de préférence avec des **applications web-dynamiques** et qui puissent permettre une **évolutivité** de la base si le recueil de données dans la base est pensé au long cours. *Par exemple (mais non-exhaustif) :* ▪ *EpiData : logiciel gratuit, ne nécessite pas de connaissances a priori, mais qui n'est pas adapté aux saisies multicentriques (en effet, ses limites sont les suivantes : pas d'accès sécurisé avec mot de passe ; données stockées sur l'ordinateur sur lequel le masque de saisie est installé ; pas d'import de données d'une source externe possible),* ▪ *REDCap : logiciel open source, intuitif ("générateur de BDD") avec des applications web-dynamiques.*

	Pour pouvoir faire du contrôle qualité sur les données, il est préférable que **l'outil de recueil soit interopérable avec l'outil de contrôle** (requête) ou bien que celui-ci soit directement intégré à l'outil de recueil.[7]
Mode de saisie	Il existe différents modes de saisie : • Recueil sur CRF papier avec saisie informatique secondaire. Bien que désuète, cette méthode ne doit pas être exclue si elle garantit une meilleure qualité. • Sur un eCRF en mode on-line, avec un objet nomade ou un ordinateur de bureau qui est alors connecté au serveur par internet. • Sur un eCRF en mode off-line sur des objets non connectés au moment de la saisie. Les données seront transférées ultérieurement lors d'une connexion au serveur. Dans ce cas, porter attention au mode et à la fréquence de synchronisation avec le serveur et également à la sécurisation du support de saisie.
Respect de la traçabilité	Toute modification de la base doit être tracée (architecture, variables, thesaurus, etc…). L'outil de recueil doit permettre de **tracer les modifications apportées aux données** (audit trail). Certains logiciels le font eux-mêmes.
Pour aller plus loin	◆ Pour le recueil de données issues d'appareils de mesure (par exemple des données de biologie) si possible préférer une exportation des données de l'appareil avec un **import direct dans la base** (sans nouvelle saisie). ◆ Afin d'anticiper l'étape de vérification des données, il est préférable d'utiliser un outil de recueil qui permet d'**alerter directement au moment de la saisie l'entrée de données aberrantes ou incohérentes.** ◆ Disposer d'une **alerte**/un signal (par exemple un code couleur) de **non-remplissage** des données obligatoires/prioritaires est un avantage pour orienter la saisie des données.

d. Les variables

Description des variables	Le **format** des variables (nombre, classe, texte, date, etc…) et leurs **unités** doivent être définis à l'avance. Si besoin la **méthode de dosage ou le moment du prélèvement** peut aussi être précisé.
Unités	**Unité de mesure usuelle** : l'unité doit être la plus proche possible de celle de la donnée-source recueillie ou sinon utiliser l'Unité Internationale.
Données non disponibles	Toujours intégrer dans l'outil la capacité de distinguer une **« donnée non disponible »** d'une donnée manquante. Bien différencier la valeur d'une variable du codage d'une donnée non saisie, non disponible ou non applicable.
Format	**Éviter au maximum les saisies libres, pour des variables non-formatées.** **Variable qualitative** : • Préférer l'utilisation de scores, d'échelle, de thesaurus existants (comme des critères diagnostics internationalement reconnus, les Dénominations

[7] Les spécificités liées à la randomisation ne sont pas traitées dans ce document (s'adresser aux Plateformes d'essais cliniques de F-CRIN).

	Communes Internationales des médicaments, le MedDRA, le CIM-10, l'ATC, etc…), • Utiliser des menus déroulants ou autres pour fixer les réponses, • Eviter le verbatim autant que faire se peut, • Prévoir et tracer l'interprétation des saisies libres avant le codage (procédure d'adjudication), • Si des zones de commentaires sont indispensables, documenter l'analyse et sensibiliser les personnes à même de les remplir sur les informations pertinentes à y faire figurer et sur le fait qu'aucune information identifiante ne doit être saisie. **Variable quantitative** : • Pour les nombres décimaux, fixer un opérateur de séparation de décimale. **Variable date** : • Préciser le format choisi et préférer un format unique pour l'ensemble de la base.
Traçabilité des modifications	Penser à tracer les transformations faites sur les variables entre la base de saisie et la base fournie pour l'analyse. L'utilisation d'une approche programmatique (suite d'instructions) est à cet égard toujours préférable aux approches de type « click bouton ».

e. Le contrôle qualité

Le contrôle qualité s'effectue sur les données saisies et doit être pensé au préalable. On distingue le monitorage centralisé ou "remote monitoring", qui est fait sur la base de données par un outil de requête (qui permet, à tout moment, de connaître l'état du recrutement et du suivi, les données manquantes, les données aberrantes, etc…), du monitoring sur site qui confronte les données saisies aux données-source en particulier.

Le niveau de contrôle et de management dépend beaucoup des enjeux de la base et de son utilisation (décision médicale, vigilance, suivi médical, santé publique, faisabilité d'essais, screening des patients, etc…).

Prévoir les contrôles suivants :

Données manquantes	Bien définir en amont quelles sont les **données obligatoires** qui ne devront donc pas être manquantes ou bien justifier si tel est le cas.
Contrôle des formats	Idéalement, le format devrait être défini dans le masque de saisie de l'outil de recueil. Le format des dates, heures et variables quantitatives est à contrôler.
Recherche de doublons	Définir les variables identifiantes pour pouvoir définir un doublon (exemple : date de naissance/âge, date de chirurgie, acte, etc…).
Données aberrantes	Définir et documenter les bornes pour considérer une valeur comme aberrante. C'est-à-dire, définir à priori un minimum et un maximum pour chaque variable, au-delà desquels la valeur doit être vérifiée.
Données incohérentes	Données incohérentes par rapport à la pathologie, entre elles ou illogiques etc…
Description de la qualité de la base	Description de l'état de la base facilement faisable sur les principaux indicateurs qualité (avec reporting, par exemple retour aux centres). *Quelques exemples d'indicateurs qualité :* • *Pourcentage de données manquantes,* • *Pourcentage de données potentiellement erronées,* • *Délai jusqu'à la correction des données erronées.*

f. Analyse statistique

Figer/Geler une base de données pour une analyse	**Identifier et dater clairement la version de la base figée/gelée** pour rendre l'analyse reproductible. Fournir par exemple **le DQM** (programme écrit) utilisé pour créer la base pour l'analyse (base datée et documentée). On peut potentiellement distinguer trois états de la base : • Base de données brute (extraction directe de la base de données proprement dite) • Base avec variables transformées (contient par exemple des variables avec des dates calendaires au lieu de 3 variables séparées pour jour, mois et année, respectivement). • Base prête pour l'analyse (contient par exemple des variables créées pour l'analyse statistique qui peuvent nécessiter des algorithmes plus complexes, comme le calcul pour un critère de jugement composite). **Toute transformation de la base brute doit être tracée.**
Analyse statistique	Les critères qualité de l'analyse statistique ne rentrent pas dans le périmètre de ce document mais les analyses statistiques doivent être faites en collaboration avec un **professionnel (statisticien ou ingénieur).**
Pour aller plus loin	Il est préférable d'utiliser un **accès à distance et direct sur la base** pour faire les analyses en évitant de dupliquer les bases (pas de transfert de données mais accès aux données). Sinon les utilisateurs ont **l'obligation de supprimer/détruire la base pseudonymisée à la fin de l'utilisation** (dans le cadre du RGPD et des CDASH).

5. Base de données intégrant des données d'imagerie et omiques

Dans ce chapitre sont développés les éléments spécifiques à ces catégories de données quand elles sont collectées en plus des données cliniques pour alimenter une base de données à vocation de recherche. Ces recommandations s'ajoutent donc à celles évoquées dans les autres chapitres du guide.

Dans le cas où ces données spécifiques seraient stockées dans une base différente des autres données (cliniques par exemple), il est essentiel d'établir des liens robustes entre les données de ces différentes bases. Dans tous les cas, il est préconisé, dans la mesure du possible, un **stockage centralisé** de ces différentes données.

a. Données d'imagerie

Il existe des plateformes spécialisées (par exemple le Centre d'Acquisition et de Traitement des Images (CATI) pour la neuroimagerie) dans le traitement et la collecte des données d'imagerie à vocation de recherche, sur lesquelles il est fortement recommandé de s'appuyer pour réaliser ce type de collecte. De telles plateformes s'avèrent indispensables dans le cadre d'une étude multicentrique à grande échelle.

Choix des données collectées	Le choix des données d'imagerie à collecter dépend de plusieurs paramètres : • **Les objectifs scientifiques :** Il est nécessaire de fixer en amont les séquences et les types d'images à collecter qui répondront aux objectifs scientifiques de la recherche. • **Le temps disponible pour réaliser les acquisitions :** Il est déterminant notamment dans les études en soin courant. En effet dans ce contexte, la nature des séquences d'acquisition sera contrainte par le temps (limité et généralement court), et leur financement, limitant les possibilités. • **Equipement des centres** : Si les données sont collectées dans plusieurs centres d'imagerie, les séquences et le type d'image seront choisis pour que tous les centres puissent les réaliser et ainsi assurer la comparabilité des

	résultats. **Le choix des données collectées dépendra donc des équipements des centres.** Dans ce type de base, il faudra prendre en considération la balance « nombre de données d'imagerie versus niveau du matériel utilisé (type de technologie et type de séquence) ».
Encadrer l'acquisition	Il est nécessaire de préciser dans un document les **détails précis sur la procédure d'acquisition** des séquences afin que les données d'imagerie puissent être comparables entre elles (données homogènes). Il est aussi conseillé de faire une **étude de faisabilité** en amont sur les centres pressentis, en sélectionnant ces centres à l'aide d'un questionnaire qualifiant selon des critères définis (avec ou sans *dummy run* sur volontaire). Il est possible d'intégrer dans la procédure un **contrôle fantôme spécifique**, en précisant sa fréquence et son but précis. *NB : Dans ce type de collecte de données on est dépendant de la gestion des parcs des machines des différents centres qui parfois peut considérablement perturber le projet de recherche.*
Collecte des images	Un monitorage précis et régulier des acquisitions est indispensable pour s'assurer que tous les examens ont bien été réalisés et transmis à la structure en charge de la gestion des images. Préférer une collecte et un contrôle qualité des images **au fil de l'eau et fréquents** pour éviter la perte de données (particulièrement si l'acquisition est faite en soin courant) et pour pouvoir refaire des examens (rescan) si la qualité des données n'était pas suffisante.
Métadonnées	Il est nécessaire de collecter en plus des images les métadonnées suivantes : les paramètres, la date et l'heure d'acquisition, les artefacts machines, le respect ou non des procédures d'acquisition. En effet, ces métadonnées permettront également de réaliser la **vérification de la qualité des données.**
Sources de variabilité	La **variabilité inter-centre** peut avoir un impact sur les analyses faites avec des données d'imagerie. **L'effet machine** : on retrouve surtout des variabilités entre différents constructeurs. Néanmoins, il faut garder à l'esprit que les variabilités inter-machine, pour 2 modèles semblables de machine, sont moindres par rapport aux variabilités de la pathologie. Un **contrôle qualité au niveau populationnel** peut s'avérer utile pour détecter des résultats suspects : centre « outlier », déviation dans le temps, changement de version du logiciel d'acquisition.

b. Données omiques

Dans le cas de base de données incluant en plus des données cliniques, des données omiques, des éléments supplémentaires sont à considérer.

Les recommandations formulées ici concernent les données omiques (c'est-à-dire des données obtenues par des techniques à haut débit permettant une analyse simultanée d'un grand nombre de variables) relatives à l'expression des gènes (exemple : les données de transcriptome (ARNm) et miRnome (microRNA)), les données du protéome, du métabolome ou du lipidome.

D'autre part, les données génétiques (génomiques) sont par définition des données identifiantes et donc des données sensibles qui répondent à des exigences réglementaires spécifiques qui ne seront pas abordées dans ce chapitre.

Techniques d'obtention des données omiques	Les techniques utilisées pour obtenir des données omiques conditionnent la façon de les traiter et de les analyser. On ne pourra pas analyser la même chose avec des données issues d'une puce à ADN, de RNA-Seq ou de PCR en temps réel (données de transcriptome), ou des données de protéomique, métabolomique ou lipidomique. Il faudra donc bien choisir le type de technique d'analyse biologique que l'on utilisera en fonction du "niveau" du vivent que l'on veut étudier. Si possible, préférer un **laboratoire centralisé** afin d'éviter les effets batchs trop importants.
Métadonnées	Il est important de recueillir et de stocker les métadonnées associées à la mesure des données omiques (exemple : appareil de mesure, etc…).
Standardisation et traçabilité	Standardiser les procédures de prétraitement et de conservation des échantillons est également essentiel, surtout en multicentrique. Leur traçabilité est également importante : délai avant pré-traitement, avant congélation, etc… Un contrôle qualité des échantillons biologiques est aussi à prévoir avant le traitement analytique.
Normalisation / prétraitement	Comme les sources et les techniques utilisées classiquement pour obtenir des données omiques sont souvent différentes d'un laboratoire à un autre, il est indispensable de normaliser ces données après la collecte et la vérification de leur qualité. Un **référentiel** doit donc être établi précisant les unités, les méthodes de prélèvement, les algorithmes de calculs, normes, etc... choisis pour chaque variable omique. D'autre part, des échantillons peuvent être déviants/divergents et il est important de les identifier et de définir leur traitement dans l'analyse statistique. Cette normalisation est chronophage et doit être tracée afin qu'elle puisse être reproductible.
Effet batch ou biais expérimental	Il est indispensable de **définir** et d'**évaluer** l'effet batch des données omiques et de le **corriger** à l'aide d'outil de bio-informatique lors du prétraitement ou de l'analyse statistique.
Stockage des données	Il est conseillé de stocker les **données brutes** (pour pouvoir par exemple appliquer un autre pré-traitement à ces données), **les métadonnées** (essentielles pour pouvoir choisir les méthodes d'analyse statistique des données et pour corriger l'effet batch) et **les données normalisées/prétraitées** (pour pouvoir refaire des analyses statistiques sur les données déjà prétraitées ou pour les partager).

6. Données sociales et de santé perçue

Pour ces catégories particulières de données, il est important de respecter quelques règles spécifiques pour assurer la pertinence du recueil.

a. Données sociodémographiques, caractérisation de l'environnement social de la personne

Collecte et choix des outils	Penser qu'il peut être utile de comparer ces données à celles de la population générale. Il est donc préférable d'utiliser **des indicateurs issus des données du recensement** (statistiques de référence à l'échelle nationale, départementale, communale), pour ce faire se référer à l'**INSEE**.

Niveau d'études	Le **diplôme le plus élevé** obtenu est l'indicateur le plus pertinent. La classification du recensement comporte 11 catégories qui peuvent être regroupées.

Situation familiale	Statut matrimonial légal selon le recensement. Pour définir ce statut, il est recommandé de se référer à la classification utilisée par INSEE :
	• Ménages composés uniquement
	  ○ D'un homme seul
	  ○ D'une femme seule
	  ○ D'un couple sans enfant
	  ○ D'un couple avec enfant(s) dont avec enfant(s) de moins de 18 ans
	  ○ D'une famille monoparentale dont avec enfant(s) de moins de 18 ans
	• Ménages complexes
	  ○ Ensemble des ménages complexes dont avec enfant(s) de moins de 18 ans

Profession	Catégorie socioprofessionnelle (PCS) : la **Classification de 2003 de l'INSEE** est la plus utilisée (en 6 catégories). La nomenclature totale inclut plus de 400 professions et est accessible en ligne s'il est nécessaire de recoder un champ libre. Un regroupement en 3 catégories peut être opéré pour distinguer les extrêmes de l'échelle sociale (ouvriers vs cadres).
Index de désavantage social	Un index de désavantage (aussi connu sous le terme anglais *deprivation*) peut être calculé à partir du lieu de résidence. Les index les plus connus sont le Townsend, le Carstairs (pour les pays anglophones), le FDep (index Français) et l'EDI (index européen). Pour avoir une donnée de qualité, il faut calculer l'index à partir de **l'adresse exacte** (incluant le n° de rue) car la seule commune de résidence donne une information trop peu précise pour les grands ensembles urbains. En effet, ces indicateurs agrègent des données sociales recueillies à l'échelle des « IRIS » (en France) : ils sont écologiques et non individuels.
Mode de vie	**Statut tabagique :** fumeur quotidien / occasionnel / non-fumeur ; année de l'initiation du tabagisme quotidien le cas échéant ; quantité et durée, pouvant être regroupées en nombre de Paquet-Année. **Alcool** : le questionnaire **AUDIT-C**, version abrégée de 3 items du questionnaire AUDIT (Développé par l'OMS) est le plus pertinent pour évaluer la consommation d'alcool. Validé dans de nombreuses populations et traduit en plusieurs langues, il inclut : les fréquences d'usage sur l'année, la quantité moyenne bue un jour standard et la fréquence de consommation d'au moins 6 verres en une occasion. D'autre part, il est utilisé dans le Baromètre Santé. **Alimentation** : il n'existe pas de questionnaire standardisé et la référence est le journal alimentaire mais ce questionnaire ne concilie pas les 3 principes de rapidité, simplicité et fiabilité. D'autre part, il n'existe pas de données en population générale sur les questions d'alimentation (à l'inverse des données sociodémographiques de l'INSEE). Ainsi, tout questionnaire ne pourra être utilisé qu'à des fins descriptives ; une comparaison nécessiterait d'établir son propre groupe de comparaison à chaque étude (utilisant un questionnaire et une méthode de mesure identiques).
Réglementaire	**L'adresse est une donnée hautement identifiante.** Pour la CNIL, toute donnée sur le mode de vie ou l'environnement social est définie comme une donnée personnelle qui nécessite une autorisation de traitement.

b. Questionnaires : Patient Reported Outcomes (PROs) / Qualité de vie

Choix des questionnaires	Critères de choix d'un questionnaire : • **Questionnaire validé (existant) ou à développer.** • **Questionnaire en accès libre ou payant.** • **Questionnaire générique** (par ex SF12, le plus utilisé) **ou spécifique d'une pathologie** : le choix d'un questionnaire générique permet des comparaisons à la population générale ou à d'autres pathologies, le choix d'un questionnaire spécifique permettra d'obtenir des données fines. • **Capacité d'utilisation** : la mesure la plus courte, la plus simple et la plus complète.
Qualité d'un questionnaire / Validation	Un questionnaire est validé par l'évaluation de ses qualités métrologiques réalisées selon des méthodes statistiques dédiées. La validation d'un questionnaire ne vaut que pour la version, la langue et la population dans lesquelles l'évaluation a été réalisée. L'ordre des items et les mots ne doivent pas être changés. Qualités métrologiques : • **Fiabilité** : la mesure est reproductible, • **Validité** : le questionnaire mesure effectivement ce qu'il est censé mesurer, • **Sensibilité au changement** : capacité à traduire l'évolution dans le temps. **L'emploi de questionnaires non validés peut conduire à des conclusions erronées.** *Attention, validité ne signifie pas toujours pertinence*.
Identification des questionnaires	2 sources : • **SeleQT,** une plateforme française développée par une équipe de chercheurs recensant les questionnaires validés (non exhaustif) et mettant gratuitement à disposition les outils nécessaires à leur exploitation https://seleqt.univ-lorraine.fr/ • Recherche de publications de validation française de questionnaire sur les moteurs de recherche de littérature scientifique.
Finalité du questionnaire	Si l'objectif est de comparer les résultats du questionnaire à la population générale ou à des populations de malades atteints d'autres maladies, il faut préférer un questionnaire générique. Si l'objectif est de réaliser des études médico-économiques de type coût-utilité, il faut préférer un questionnaire permettant le calcul d'utilité (ex : EQ5D).
Méthode de recueil des données	Le **mode de passation** (auto-questionnaire ou entretien, seul ou avec une aide, papier ou informatique) influence la qualité et la nature des réponses (nombre de données manquantes, biais de déclaration). Ces modalités sont à prendre en compte lors du choix de la méthode de recueil et à renseigner lors de la saisie des données (**métadonnées**). D'autre part, la méthode de recueil choisie devra être documentée en amont de tout recueil pour assurer l'homogénéité des résultats.
Interprétation	Certains questionnaires ont des seuils de référence (ex : patient dépressif si score supérieur à une valeur donnée), d'autres non (ex : score de fatigue). Dans le dernier cas, la différence cliniquement pertinente sera à identifier dans la population d'étude pour interpréter des données.
Data management et analyse	Les questionnaires nécessitent un calcul de score, le scoring et la pondération de chaque item. Ces éléments sont détaillés sur le site internet seleQT ou dans les articles de construction ou de validation des questionnaires. Prévoir des méthodes de réduction du risque de données manquantes lors du recueil et lors de leur analyse.

Enfin, pour des projets de recherche clinique (hors recherche en Sciences Humaines et Sociales, très spécifique), en fonction de l'objectif de la recherche, le recueil de chaque item des questionnaires est souvent non nécessaire. On préfèrera dans ces cas recueillir les scores totaux ou sous-scores de ces questionnaires *(cf. chapitres "4.b. Aspects déterminants dès la conception du projet" et "4.d. Les variables")*.

Attention lors de l'analyse à ne pas commettre **l'erreur écologique** (« *ecological fallacy* ») et à envisager **l'utilisation de modèles multiniveaux.** Il est donc nécessaire de collaborer avec des **experts de l'analyse de cette catégorie de données.**

7. Aspects Juridiques et Règlementaires

a. RGPD : les bons réflexes

a. 1. Démarches à effectuer : AIPD, MR, demande d'autorisation de traitement…

✓ **AIPD**

Pour toute création d'une base de données de recherche en santé, il est nécessaire de vérifier assidument si une **Analyse d'Impact relative à la Protection des Données** (ou PIA "Personnal Impact Assessment")[8] doit être réalisée ou non. Cette analyse doit être entreprise dans la majorité des projets de base de données de recherche en santé, compte tenu des risques que présentent ces bases pour les droits et libertés des personnes concernées. Pour aider dans cette démarche, la CNIL fourni un logigramme de prise de décision dans ce sens[9] et elle a listé les différents types d'opérations de traitement pour lesquelles une analyse d'impact relative à la protection des données est requise[10].

L'analyse d'impact doit être mise à jour après mise en application des mesures ciblées dans le plan d'action ou si un nouveau type de traitement est effectué.

L'analyse d'impact pourra être demandée par la CNIL dans le cadre de l'instruction de la demande d'autorisation.

Elle sera systématiquement requise et doit être jointe à la demande d'autorisation en cas de :

- Recherche médicale incluant un traitement portant sur des données génétiques de patients et/ou de personnes mineurs,
- Constitution d'un registre ou d'une base de données ayant vocation à être mise à la disposition des communautés de recherche.

Une AIPD faisant apparaître des risques résiduels élevés malgré les mesures envisagées par le responsable de traitement concerné doit être transmise à la CNIL à l'appui d'une demande d'autorisation.

Pour les recherches relevant d'une Méthodologie de Référence, l'analyse d'impact doit être réalisée et tenue à la disposition de la CNIL mais n'a pas à lui être adressée. Elle sera demandée en cas de contrôle.

[8] Voir *"2. Principales définitions et abréviations"* et pour en savoir plus et connaître les formalités applicables : https://intranet.inserm.fr/securite-et-prevention/protection-donnees-personnelles/reglementation-generale/Pages/formalites.aspx En cas de difficulté, contacter le DPO de l'Inserm (dpo@inserm.fr) ou un des membres du réseau référent à la protection des données.

[9] https://www.cnil.fr/fr/infographie-dois-je-faire-une-aipd

[10] https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-avec-aipd-requise-v2.pdf

✓ **MR ou Autorisation de traitement**

La possibilité d'utiliser une Méthodologie de Référence particulière dépend du type de projet et de données que l'on envisage d'utiliser. Dans ce but, l'INDS à mis à disposition des modes d'emploi de ces MR à destination des professionnels de la recherche.[11]

Enfin, si le projet ne répond pas à une Méthodologie de Référence il faudra compléter et envoyer une "Demande d'autorisation d'un traitement de recherche dans le domaine de la santé"[12]. Le circuit de circulation des données dépendra de la qualification réglementaire de la recherche selon qu'elle implique ou non la personne humaine.

a. 2. Consentement et Notice d'Information

Pour savoir si l'utilisation d'un consentement est nécessaire dans un projet de base de données, se référer aux recommandations de classification des projets de recherche clinique (non-RIPH, RIPH1, RIPH2 ou RIPH3).

Le consentement et la notice d'information associée doivent explicitement prévoir les éléments suivants (pour plus d'informations, se reporter au paragraphe ***"4.a Prérequis _ Consentement et Notice d'information : « partage et réutilisation des données »"***) :

Moyens et modalités d'exercice des droits	**Les droits du participant au titre du RGPD et les moyens efficaces à sa disposition pour les exercer** auprès du responsable du traitement ou de son représentant. Fournir des indications claires et des étapes simples. Les personnes doivent être clairement informées des **modalités pratiques** d'exercice des droits, et les démarches à effectuer ne doivent pas décourager les personnes concernées ou occasionner des frais. **Le droit d'introduire une réclamation auprès d'une autorité de contrôle.**
Coordonnées des personnes référentes	**Les coordonnées du délégué à la protection des données**. Inviter les personnes concernées à contacter en premier lieu la personne les ayant pris en charge (l'investigateur dans le cas de RIPH), qui connaît leur identité et pourra permettre l'exercice effectif des droits, ce qui n'est pas le cas du DPO. Diriger les ensuite vers le DPO en cas de difficultés. **L'identité et les coordonnées du responsable du traitement** et, le cas échéant, du représentant du responsable du traitement.
Accès aux données	**Le caractère international des accès aux données** si applicable (ex : avec des partenaires ou des catégories de partenaires (à préciser) établis dans des pays membres de l'Union Européenne ou dans des territoires assurant un niveau adéquat de protection des données personnelles). **Accès ouvert aux données** (à tout public ; données anonymes) **ou restreint** (sur contrôle des demandes d'accès). *Attention, le libre accès n'inclut pas par défaut une libre réutilisation des informations à caractère personnel. Les réutilisations peuvent être soumises à des conditions spécifiques (ex : examen préalable des demandes d'accès et exigences de probité éthique et règlementaire de l'accédant en fonction des lois applicables à son activité) et à un des régimes de propriété intellectuelle existant*.
Destinataires des données	**Les destinataires prévus des données collectées** (partenaires ou des catégories de partenaires, à préciser).

[11] https://www.indsante.fr/fr/actualite/publication-des-nouvelles-methodologies-de-reference-mr
[12] https://www.service-public.fr/professionnels-entreprises/vosdroits/R18457

Partage des données	**Le partage des données**, ses raisons, avec quels types de structure (public et/ou privé) et son caractère obligatoire ou non. **La nature des données partagées** : partage sous une forme anonymisée ou pseudonymisée dans le respect du secret médical. **La base légale du partage**.
Réutilisation des données	**La réutilisation des données et les finalités des réutilisations des données** : délimitation claire, plus ou moins large, du périmètre des réutilisations (exemple de délimitation large : domaines d'études, groupes de pathologies).
Fusion/chaînage	**La fusion / le chaînage avec une ou des bases médico-administratives** (celles du SNDS par exemple).

b. Transfert de données

Avant tout projet de transfert, il faut bien avoir à l'esprit qu'il s'agit de données personnelles partiellement ou complètement identifiantes et soumises, outre à la réglementation, à des droits de propriété intellectuelle.

D'autre part, toute demande d'accès à des données personnelles (directement ou indirectement identifiantes) doit donner lieu à une **évaluation basée sur des critères éthiques et juridique pertinents** pour le domaine d'activité concerné. Ceci inclut la **vérification de la compatibilité du consentement et de l'information** initialement donnée à la personne concernée avec les utilisations envisagées dans la demande d'accès.

Conditions préalables	Dans le cadre d'un **protocole d'accord défini** (cf. accord de consortium). Le transfert se fait sur la base figée à un temps précis et défini. Il est primordial de respecter le **principe de minimisation** des données par rapport aux nécessités liées à la finalité d'utilisation des données demandées. Attention : Le **transfert d'échantillons biologiques d'origine humaine à l'étranger** (export/import) nécessite une autorisation préalable spécifique de l'organisme français exportateur/importateur délivrée par le Ministère de la Recherche.
Outils	Utiliser des outils de transfert **sécurisés** choisis avec l'institution support (à ce titre, le transfert d'une base de données par mail, par Dropbox ou par Wetransfer est totalement inapproprié). Il est préférable de crypter les données avant de les transférer. La clef de décryptage étant transmise au destinataire séparément des données.
Type de données transférables	Prévoir si le transfert porte sur des **données brutes pseudonymisées** ou des **données prêtes pour l'analyse (non-identifiantes également)**. *Dans ce dernier cas il faut toujours utiliser le même algorithme de création de variables pour l'analyse, notamment pour les variables complexes et le tracer dans un DQM.* Dans la mesure du possible, transférer des **données pseudonymisées ou anonymisées** après information des participants, mais jamais des données directement identifiantes (sauf nécessité absolue qui doit être justifiée et documentée).
Traçabilité	Documenter le transfert (date, type de données, destination, durée de conservation de la base figée, etc…) et associer la copie de la base figée transférée. Veiller à encadrer le transfert par un **contrat de transfert ou data transfer agreement (DTA)** permettant de garantir que les données seront désormais sous la responsabilité du destinataire. *Il est rappelé que les données personnelles sont sous la responsabilité du responsable de traitement et que tout transfert de données à des tiers non*

	autorisés ou dans des conditions non cadrées est une violation de l'obligation de sécurité. Tout problème de sécurité des données survenu chez l'importateur des données serait sous la responsabilité de l'exportateur.
Destinataire	**Transfert vers un pays de l'UE : L'ensemble des Etats Membres de l'UE est considéré comme assurant une protection adéquate des données au sens du RGPD.** Un tel transfert ne nécessite pas d'autorisation spécifique. Néanmoins, compte tenu du caractère sensible des données de santé, des données génétiques, biométriques, des données qui révèlent la vie sexuelle ou l'orientation sexuelle d'une personne physique, la prétendue origine raciale ou l'origine ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale d'une personne physique, il est fortement recommandé de conclure un **contrat de transfert (DTA) ou tout autre accord juridiquement contraignant** destiné à encadrer l'échange et les utilisations possibles des données. **Transfert vers un pays tiers à l'UE** : Ce type de transfert répond à des **conditions strictes et arbitraires**. Dans ce cas, il est nécessaire de se rapprocher du promoteur de la base. Tout responsable de traitement qui souhaite exporter des données à caractère personnel hors de l'UE doit : • Se renseigner sur le niveau de protection adéquat du pays destinataire[13], • Informer les personnes participantes des transferts envisagés, faute d'adéquation, du recours à des « garanties appropriées », et indiquer le moyen d'obtenir une copie des données ou l'endroit où elles ont été mises à disposition. *Il est conseillé de prévoir des conditions élargies même si cela peut sembler alourdir le consentement car revenir en arrière est difficile voire infaisable en pratique.*
Utilisation des données	L'utilisation des données et donc leur transfert doivent être faits et pensés pour une question précise, définie au préalable. *Tout transfert de données en vue d'une nouvelle recherche donnera lieu à un nouveau traitement qui sera soumis à l'ensemble des principes de protection des données personnelles (finalité, licéité, minimisation des données, conservation limitée, sécurité) et à des autorisations réglementaires s'il y a lieu.*
Pour aller plus loin	Il est préférable d'utiliser un **accès à distance et direct sur la base** pour faire les analyses en évitant de dupliquer les bases (pas de transfert de données mais accès aux données). Sinon les utilisateurs ont **l'obligation de supprimer/détruire la base pseudonymisée à la fin de l'utilisation** (dans le cadre du RGPD et des CDASH).

Enfin, par principe, les données à caractère personnel de santé ne peuvent être vendues, cédées ou exploitées commercialement. En revanche les résultats de l'exploitation de ces données le peuvent (en d'autre terme, c'est l'usage et les services autour des données qui peuvent être valorisés commercialement).

[13] Faute de niveau de protection adéquat, le transfert reste néanmoins possible moyennant des garanties supplémentaires y compris lorsque ces données sont codées ou chiffrées.

c. Spécificités relatives à une Biobanque

En règle générale, les biobanques ont une base de données associée, utile pour la gestion des échantillons biologiques et les critères de qualité listés plus haut s'appliquent donc aussi à ces bases.

Démarches réglementaires	Formalités à réaliser liées à l'activité de préparation, conservation et gestion de la collection d'échantillons humains pour la biobanque : • **RIPH : Avis favorable CPP et Autorisation ANSM pour RIPH1,** • **Hors RIPH : Procédures CODECOH avec le Ministère de la Recherche,** • **Autorisation de transfert d'échantillons biologiques humains,** • **RGPD** à respecter dans tous les cas.
Référentiels normatifs	Certification NF S 96-900 (France) ou ISO 20387:2018 (International).
Respect du RGPD	Respecter une gestion éthique des dépôts et des accès aux données et des échantillons de la biobanque, en définissant des mesures organisationnelles et techniques assurant la confidentialité, le respect des consentements et l'exercice des droits des personnes (y compris en cas de retrait du consentement ou de désengagement).
Information des participants	Procédure claire et accessible au public concernant les activités (lieu de stockage, transfert, analyse, etc…) de la biobanque et le devenir des données/échantillons en cas de fermeture (Politique interne, plan de gestion des données; Evaluation d'Impact sur la Vie Privée régulièrement mise à jour).
Principes FAIR	Respecter les principes FAIR au regard des collections d'échantillons et des données associées. Des lignes directrices sont disponibles en anglais[14].
Autorisation de transfert	Le transfert d'échantillons biologiques (export/import) nécessite une autorisation préalable spécifique délivrée par le Ministère de la Recherche à l'organisme français exportateur/importateur.

d. Accords de consortium, Aspects juridiques

Ce chapitre traite des aspects juridiques relatifs à l'accès aux données et à leur utilisation, au partage des données et à la valorisation de la base. Les aspects juridiques sont à **anticiper**, autant que faire se peut, en amont de la construction de la base car ils conditionnent certains aspects de l'utilisation des données et parce que la discussion est toujours plus fluide avant l'obtention de résultats et plus conflictuelle ou tendue si ces aspects sont à gérer au moment de la valorisation des travaux.

Ci-dessous, une liste non exhaustive d'éléments à ne pas omettre, pour faciliter l'établissement des documents nécessaires, comme l'accord de consortium, le data transfer agreement, le contrat d'hébergement, etc… Ces documents doivent fixer les éléments suivants :

✓ **Les accès et l'utilisation aux données**

Il faut distinguer :

1- Le périmètre d'accès aux données : toutes les données, accès aux données d'un centre, accès à une catégorie de données, etc…
2- Le format des données : données individuelles, données agrégées, métadonnées (qui décrivent les données de la base)
3- Les droits d'accès et d'utilisation :
   Aux membres du consortium,

[14] https://www.go-fair.org/fair-principles/

Aux scientifiques académiques,
À des compagnies privées.
Il est possible d'avoir un modèle en "open access" avec des conditions d'utilisation des données.

- ✓ **La valorisation**

1- Valorisation scientifique avec les règles de publication : penser, en particulier, à définir la préférence entre un open access green et/ou gold (encouragé) :

- La voie verte (ou green open access)[15] est la voie de l'auto-archivage ou dépôt par l'auteur dans une archive ouverte. Une archive ouverte est un réservoir où sont déposées des publications issues de la recherche scientifique et de l'enseignement dont l'accès est libre et gratuit.
- La voie dorée (ou gold open access) concerne des revues ou ouvrages nativement en open access, dès leur publication et qui ont fait l'objet d'un paiement préalable à leur mise en open access permettant de couvrir les coûts d'édition.

2- La valorisation financière :

La facturation de l'exploitation de la base est différenciée selon le demandeur. Elle est fonction de la nature de la demande et de l'importance de la requête (fonction de la nature/des variables de données, du nombre, de la période de traitement des données mobilisées).
Assez classiquement, mais pas obligatoirement :

- Les accès et les utilisations sont gratuits pour les membres du consortium,
- Les accès et les utilisations pour les académiques sont facturés au prix coûtant de la préparation de la base et du transfert (data management et autre),
- Les accès et les utilisations pour des industriels sont facturés et permettent ainsi la valorisation de la base.

- ✓ La **gouvernance scientifique** : par exemple, organigramme des comités de l'étude et définition de leurs missions.
- ✓ La **gouvernance juridique et administrative** reflétée dans **l'accord de consortium**, fixe le partage des responsabilités et les missions, droits et obligations de chacun des acteurs (coordonnateur, responsable scientifique, centre, conseil de gouvernance, conseil scientifique par exemple).
- ✓ L'accord de consortium régissant la gouvernance de la base permet de définir **la propriété intellectuelle** qui conditionne l'accès et l'utilisation/l'exploitation des données et, en particulier, le partage des "royalties" éventuelles entre les membres du consortium.
- ✓ L'obligation de **traçabilité** de tout changement de la structure de la base dans un document.
- ✓ **L'hébergement** de la base : la prestation d'hébergement des données personnelles de santé nécessite un **contrat spécifique** (conforme à l'article 28 du RGPD) devant prévoir que les hébergeurs ne peuvent utiliser les données qui leur sont confiées à d'autres fins que l'exécution de la prestation d'hébergement, la restitution des données à la fin de l'hébergement et le respect du secret professionnel.
- ✓ Le **transfert** de données dans le cadre de "prestation de service en dehors d'un accord de consortium" : utilisation de contrats dédiés (**Data Transfer Agreement**).
- ✓ Les **conditions financières :** l'établissement d'une base de données et sa maintenance ont un coût qu'il faut prévoir et assurer.
- ✓ La gestion de la **confidentialité.**

[15] https://openaccess.couperin.org/la-voie-verte-2/

8. Références juridiques et réglementaires

Liste (non exhaustive) des références juridiques/réglementaires en vigueur lors de la rédaction de ce guide (avec les liens hypertextes) ; Attention ce document a été construit sur la base des versions en vigueur à la date de la rédaction initiale de ce guide, le 24/09/2019 :

- AIPD : Liste des types d'opérations de traitement pour lesquelles une analyse d'impact relative à la protection des données est requise, logigramme dois-je faire une AIPD ? et Ce qu'il faut savoir sur l'analyse d'impact relative à la protection des données (AIPD)
- Association Médicale Mondiale :
  - Déclaration de Taipei sur les bases de données de santé et les biobanques.
  - Déclaration d'Helsinki – Principes éthiques pour la recherche médicale impliquant des sujets humains
- CCNE, Avis 130 sur « Données massives (big data) et santé : une nouvelle approche des enjeux éthiques »
- Bonnes Pratiques Cliniques (Françaises, Européennes) et Avis aux promoteurs et aux investigateurs Inserm
- Clinical Data Acquisition Standards Harmonization
- Code de la santé publique
- Conseil de l'Europe: Convention 108+ et Recommandation sur la protection des données de santé
- Demande d'autorisation d'un traitement de recherche dans le domaine de la santé
- EMA ICH E18 on genomic sampling and management of genomic data
- ISO/TC 276: Biotechnologie et biobanques
- ISO 9001: Système de management de la qualité
- ISO 20000: Système de gestion de la qualité des services
- ISO 27001: Système de management de l'information
- OCDE, Lignes directrices sur les biobanques et bases de données de recherche en génétique humaine.
- Loi Jardé, Ordonnance et Arrêtés fixant la liste des recherches RIPH 2 et RIPH 3
- Modes d'emploi des Méthodologies de Référence
- Procédures CODECOH
- Référentiels de certification des Hébergeurs de Données de Santé (Agence française de la santé numérique - ASIP Santé)
- Règlement (UE) Essais Cliniques
- RGPD, Lignes directrices du CEPD et Loi Informatique et Libertés

9. Les contributeurs

Ce guide a pu voir le jour grâce à une collaboration effective et une forte implication de différents spécialistes de leur domaine d'activité. La diversité des thématiques qui ont pu être abordées pour la création de ce guide renforce une nouvelle fois le sentiment que F-CRIN favorise la réunion et le partage des expertises de la recherche en santé présentes en France.

Ce groupe de travail sur la qualité des bases de données a permis de réunir les personnes suivantes, que nous tenons à remercier pour la qualité du travail qu'ils ont fourni et pour leur grande disponibilité :

Nom	Composante de rattachement	Sollicité au titre de
Vincent BOUTELOUP	CHU Bordeaux	Statisticien
Gauthier CHASSANG	Infrastructure BIOBANQUES et Plateforme Genotoul Societal	Juriste
Agnès DUMAS	PARTNERS	Chercheuse en sciences sociales
Hélène ESPEROU	Responsable du pôle Recherche Clinique INSERM	Observatrice
Jean-Christophe HEBERT	Responsable du Département des affaires juridique INSERM	Juriste
Christian JACQUELINET	INI-CRCT	Médecin coordinateur – Agence de la Biomédecine
Delphine JEAN	EUCLID	Data manager
Martine LAVILLE	FORCE	Coordinatrice de réseau
Enora LE ROUX	PARTNERS	
Stéphane LEHERICY	CATI	Radiologue
Frédérique LESAULNIER	INSERM	Déléguée à la protection des données
Claire LEVY-MARCHAL	Bureau Exécutif F-CRIN	Coordonnatrice du groupe de travail
Mayka MERGEAY-FABRE	F-CRIN Coordination	Secretariat scientifique du groupe de travail – Chef de projet en recherche clinique
Ivan MOSZER	NSPark	Chercheuse nutrition
Laura RICHERT	EUCLID	Coordonnatrice du groupe de travail - Epidémiologiste
Christine ROUSSEAU	FORCE	Data Manager
Sabrina VERCHERE	PARTNERS	Attachée de recherche clinique senior
Nathalie VIGUERIE	FORCE, Inserm	Chercheuse Nutrition

Nous remercions également les personnes suivantes pour leur relecture consciencieuse du guide : Clémence CAMBERLEIN (F-CRIN Coordination), Geneviève CHENE (EUCLID), Anne CLAUZEL (CRICS-TRIGGERSEP), Jean-Christophe CORVOL (NS-Park), Marie-Dominique Devignes (INI-CRCT), Stephan EHRMANN (CRICS-TRIGGERSEP), Michel PAQUES (FRCRnet), Carole PIERRART (INSERM), Margot PREVOST (RECaP), Aurélie RABIER (FRCRnet), Grégoire REY (Cépi DC, INSERM), Jacques ROPERS (RECaP), Cédrick WALLET (EUCLID).

➔ *Pour toutes suggestions d'amélioration de ce guide, veuillez les transmettre à l'adresse contact@fcrin.org.*