Modèle de protocole scientifique

A remplir

|  |
| --- |
| ***Aide au remplissage***  *(*Cette partie est à effacer avant envoi du projet*)* |

Pour prendre en compte votre demande d’autorisation d’accès à des données personnelles de santé, nous avons besoin que vous remplissiez ce protocole en respectant la structure, le format et le nombre de pages afin de faciliter l’instruction par le [CESREES](https://www.health-data-hub.fr/cesrees), puis la CNIL. Par ailleurs, la taille de la police doit être standard selon les usages en vigueur (Calibri 10 minimum par exemple). **Le non-respect de ces consignes peut justifier la non instruction de votre dossier.**

Le CESREES est un comité indépendant qui évaluera votre dossier au regard de la finalité et la méthodologie de la recherche, la nécessité du recours à des données de santé à caractère personnel, la pertinence éthique, la qualité scientifique du projet et le caractère d'intérêt public que présente le projet le cas échéant.

Par ailleurs, nous ne pouvons garantir que des pièces complémentaires à celles prévues dans la liste des pièces du dossier, que vous souhaiteriez partager en plus, soient effectivement lues.

Dans la suite du protocole, nous vous invitons à effacer systématiquement les parties “Aide au remplissage” et “Nos conseils” pour ne laisser apparaître que vos réponses.

En cas de doute sur les démarches réglementaires à accomplir pour accéder aux données de santé et éventuellement l’éligibilité de votre étude à une méthodologie de référence, vous pouvez contacter le HDH ([hdh@health-data-hub.fr](mailto:hdh@health-data-hub.fr)) ou consulter le site du [HDH](https://www.health-data-hub.fr/demarches-reglementaires) ou de la [CNIL](https://www.cnil.fr/fr/traitements-declaration-conformite?field_norme_numerotation_type_value%5B0%5D=6).

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

Si vous voulez accéder à des données dont vous n’êtes pas le dépositaire, vous aurez besoin d’aide pour remplir les rubriques “3.1 - Données requises”, “4.1 - Respect des droits des personnes concernées” et “4.2- Confidentialité et sécurité des données”. Nous vous conseillons vivement de prendre contact avec les équipes en charge de ces données. Si vous n’allez pas héberger vous-même les données nécessaires pour votre projet, vous aurez besoin de l’aide du responsable du système d’information concerné pour remplir les rubriques “4.2.2 - Support des données”.

Attention, aujourd’hui 1/3 des dossiers doivent être retravaillés après un passage devant le CESREES car les éléments soumis sont insuffisants ; y consacrer le temps nécessaire vous fera gagner du temps par la suite !

Il paraît difficile de réunir toutes les informations nécessaires au remplissage de ce dossier en moins de 2 mois. En cas de besoin, nous pouvons vous aider ici <https://www.health-data-hub.fr/contact>, ainsi que la communauté [ici](https://entraide.health-data-hub.fr/).

A toutes les étapes de la démarche, vous pouvez :

* Consulter le [répertoire des projets](https://www.health-data-hub.fr/projets) pour avoir une idée de ce qui est réalisé à partir des données
* Faire appel à la [communauté](https://entraide.health-data-hub.fr/)
* Participer aux [formations](https://health-data-hub.fr/formations) organisées par le HDH pour avoir un premier niveau d'information sur le [SNDS](https://documentation-snds.health-data-hub.fr/) et aux [meet-up](https://www.meetup.com/fr-FR/Health-Data-Hub/) pour échanger avec tout type de porteurs de projet
* Vous référer à la [documentation collaborative](https://documentation-snds.health-data-hub.fr/) ou utiliser un des programmes partagés par Ia communauté



**[TITRE COMPLET DE LA RECHERCHE]**

**Sommaire du protocole scientifique**

[**PRÉSENTATION DE L’ÉQUIPE PROJET [1 PAGE MAXIMUM]**](#_heading=h.gjdgxs) **5**

[**OBJECTIFS ET FINALITÉS [2 - 4 PAGES MAXIMUM]**](#_heading=h.30j0zll) **7**

[Contexte, objectif(s) et justification de l’étude](#_heading=h.1fob9te) 7

[Justification du respect de l’éthique](#_heading=h.3znysh7) 7

[Justification de l’intérêt public](#_heading=h.2et92p0) 7

[Publication des résultats et valorisation](#_heading=h.tyjcwt) 9

[**MÉTHODOLOGIE [3 - 5 PAGES MAXIMUM]**](#_heading=h.1t3h5sf) **10**

[Design de l’étude](#_heading=h.4d34og8) 10

[Description et justification de la population d’étude](#_heading=h.2s8eyo1) 10

[Taille de la population](#_heading=h.17dp8vu) 12

[Sources de données](#_heading=h.26in1rg) 12

[Variables](#_heading=h.1ksv4uv) 13

[Préparation de données](#_heading=h.2jxsxqh) 13

[Méthodes, traitements et analyses des données](#_heading=h.z337ya) 14

[Limites de l’étude](#_heading=h.3j2qqm3) 14

[Calendrier prévisionnel et faisabilité du projet](#_heading=h.1y810tw) 15

[**PROTECTION DE LA VIE PRIVÉE, SÉCURITÉ ET CONFIDENTIALITÉ DES DONNÉES [10 PAGES MAXIMUM] (Section destinée à la CNIL)**](#_heading=h.4i7ojhp) **16**

[Respect des droits des personnes concernées](#_heading=h.2xcytpi) 16

[Confidentialité et sécurité des données](#_heading=h.1ci93xb) 20

[Gestion du risque de réidentification](#_heading=h.3whwml4) 20

[Support de données](#_heading=h.2bn6wsx) 21

[Circuit des données en cas d’appariement](#_heading=h.qsh70q) 22

[**ANNEXES**](#_heading=h.3as4poj) **26**

# PRÉSENTATION DE L’ÉQUIPE PROJET [1 PAGE MAXIMUM]

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Désigner les structures Responsable(s) de traitement et du/des responsable(s) de mise en oeuvre du projet :

* Le / les responsable(s) de traitement sont : si besoin vous référer à la définition ci-après.
* Le / les responsable(s) de la mise en œuvre sont : si besoin vous référer à la définition ci-après.

Présentez les.

Indiquer le nom de l’organisme du financeur.

Désigner les représentants de ces structures et lister les membres de l’équipe projet, leur fonction, et leur rôle au sein du projet.

Désigner le Délégué à la Protection des Données. Nous vous conseillons de préciser son nom, son adresse postale, téléphone, mail, site web.

Si dans votre étude un comité scientifique est mis en place, préciser son rôle, mentionner ses membres et leurs affiliations.

|  |
| --- |
| ***Définition***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Le responsable d’un traitement** de données à caractère personnel est la personne, l’autorité publique, le service ou l’organisme qui détermine les finalités, les objectifs et les moyens de l’étude. Il endosse la responsabilité juridique du traitement et l’autorisation CNIL est établie à son attention. Voici une liste d’indices permettant d’identifier le responsable de traitement :

* Il est à l'initiative du traitement ;
* Il décide de l'objectif et des résultats du traitement ;
* Il décide quelles données vont être collectées ;
* Il décide quelle population va être concernée par le traitement ;
* Il a une totale autonomie dans la détermination des moyens et finalités du traitement ;
* Il choisit les sous-traitants pour traiter les données pour son propre compte ;
* Le responsable de traitement doit être représenté par son représentant légal.

**Le responsable de mise en œuvre** est en charge de réaliser l’étude ou une partie de l’étude ; s’il n’est pas identique au responsable de traitement, il est son sous-traitant.

**Le financeur** est l’organisme qui fournit tout ou partie des capitaux nécessaires à la réalisation de l’étude, recherche ou évaluation. S’il est identique au responsable de traitement, ce champ n’est pas à remplir.

**Le délégué à la protection des données** assure la qualité de l’information des personnes concernées par l’étude et s’assure de l'absence d’opposition de ces personnes, conformément au RGPD et la loi informatique et liberté (LIL).

**Le comité scientifique de l’étude** est un comité d’experts indépendants mis en place par l’équipe projet pour assurer le bon déroulement de l’étude d’un point de vue scientifique. Ce comité est généralement composé au minimum de 3 membres de différentes spécialités (par ex. un(e) neurologue, un(e) épidémiologiste et un(e) économiste de la santé). Le comité scientifique est facultatif et ne s’applique pas à toutes les études. Il est cependant préconisé pour les études nécessitant une expertise particulière ou ayant recours à des méthodologies innovantes.

**Le gestionnaire du système fils** est un l’organisme responsable du système fils. Le système fils est un système du SNDS qui rassemble des données du SNDS provenant déjà d'un système du SNDS et ii) assure leur mise à disposition dans le cadre des finalités du SNDS (voir art. L. 1461-1 du code de la santé publique).

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Dans le cas où vous êtes responsable de traitement et que vous en assurez vous-même les analyses (donc responsable de mise en œuvre)**, précisez-le. En effet, la réglementation (RGPD) en vigueur fait intervenir notamment deux types de rôles dans un projet d’utilisation de données de santé : le responsable de traitement qui est à l’initiative de l’étude ; le responsable de mise en œuvre qui est chargé de l’exécuter. Toutefois, en pratique, ces deux rôles peuvent être tenus par la même personne : par exemple, si vous êtes à l’initiative de l’étude (donc responsable de traitement) et que vous en assurez vous-même le traitement (donc responsable de mise en œuvre).

**En cas de co-responsabilité de traitement,** nous vous conseillons d’expliquer le partage des responsabilités dans le protocole scientifique, la CNIL sera attentive à ce point. Ces éléments de rôles et responsabilités attenants à chaque partie seront issus d’une convention établie en amont entre ces co-responsables.

**Pour les thèses de médecine et autres mémoires**, il convient de distinguer la personne qui réalise le travail (le / la doctorant(e)) et celle qui l’encadre (directeur/trice de thèse) et qui, sous l’autorité du responsable de traitement, est le/la responsable de la mise en œuvre.

# OBJECTIFS ET FINALITÉS [2 - 4 PAGES MAXIMUM]

## Contexte, objectif(s) et justification de l’étude

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Indiquer le contexte général de l’étude (scientifique, médical, organisationnel, épidémiologique, etc.). Préciser la pertinence du projet au regard de l'état actuel des connaissances et l'impact attendu.

Décrire les finalités de l’étude (amélioration de la prise en charge des patients, évaluation des politiques de santé, surveillance sanitaire…).

Indiquer le(s) objectif(s) de l’étude, les critères d’évaluation associés et la nature du(es) bénéfice(s) (scientifique, médical, social, ou concerne l’organisation du système de santé).

## Justification du respect de l’éthique

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Préciser en quoi l’étude est conforme aux principes éthiques, qu’elle ne pose pas de problème éthique, pas de stigmatisation d’un groupe spécifique et ne va pas à l’encontre de la morale.

## Justification de l’intérêt public

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Démontrer que le projet poursuit un intérêt public notamment en indiquant :

* Les finalités du projet : elles doivent être exposées de façon claire, intelligible et sincère ;
* Le bénéfice du projet : il doit apporter un bénéfice direct ou indirect pour les individus, pour la société, ou pour la communauté scientifique ; l’intérêt commercial n’est pas incompatible avec l’intérêt public ;
* Les raisons pour lesquelles la finalité poursuivie n’entre pas dans une finalité interdite ;
* Les efforts et engagements de transparence et de publication des résultats ;
* Les dispositions prises afin de garantir l’intégrité scientifique et la qualité des études et de prévenir le risque de produire des résultats biaisés. Ces dispositions peuvent prendre la forme d’une implication d’acteurs professionnels de la recherche, de la mise en place d’une gouvernance scientifique adaptée, ou encore d’une ouverture d’éléments de méthodes ou codes (par ex. algorithmes) permettant de discuter et vérifier la validité des résultats ;
* Pour plus d’informations, vous pouvez vous reporter au site internet du Health Data Hub : [Intérêt public](https://www.health-data-hub.fr/interet-public).

|  |
| --- |
| ***Définition***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Toutes les études, recherche et évaluation sur des données personnelles de santé doivent poursuivre une finalité d’intérêt public pour être autorisées**. C’est l’article 66 de la loi Informatique et Libertés : "Les traitements relevant de la présente section ne peuvent être mis en œuvre qu'en considération de la finalité d'intérêt public qu'ils présentent. La garantie de normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux constitue une finalité d'intérêt public.”

S’agissant du Système National des Données de Santé[[1]](#footnote-1), en plus de poursuivre une finalité d’intérêt public, les projets doivent s’inscrire dans au moins une des finalités suivantes :

* A la recherche, aux études, à l’évaluation et à l’innovation dans les domaines de la santé et de la prise en charge médico-sociale ;
* A l’information sur la santé ainsi que sur l’offre de soins, la prise en charge médico-sociale et leur qualité;
* A la définition, à la mise en œuvre et à l’évaluation des politiques de santé et de protection sociale ;
* A la connaissance des dépenses de santé, des dépenses d’assurance maladie et des dépenses médico- sociales ;
* A l’information des professionnels, des structures et des établissements de santé ou médico-sociaux sur leur activité ;
* A la surveillance, à la veille et à la sécurité sanitaires ;

Certaines finalités sont par ailleurs strictement interdites. **Il est interdit de proposer un traitement** soit qui aurait pour objectif d’aboutir à prendre une décision concernant une personne physique identifiée sur le fondement des données la concernant et figurant dans l’un de ces traitements, soit qui viserait :

* La promotion des produits de santé en direction des professionnels de santé ou des établissements ;
* Ou l’exclusion de garanties des contrats d’assurance ou la modification de cotisations ou de primes d’assurance pour un individu ou un groupe d’individus.

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Si le responsable de traitement est un industriel du produit de santé ou un assureur,** il devra :

* Soit démontrer que les modalités de mise en œuvre du traitement rendent impossible toute utilisation des données pour l'une des “finalités interdites”;
* Soit recourir à un laboratoire de recherche ou à un bureau d'études, publics ou privés, pour réaliser le traitement (en tant que responsable de la mise en œuvre). Un contrat de sous-traitance reprenant les exigences de l’article 28 du RGPD devra être conclu. Ces laboratoires de recherche ou bureaux d’études réalisant des études pour des tiers devront en outre s’engager auprès de la CNIL à respecter [le référentiel](https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000035268202/) déterminant les critères de confidentialité, d’expertise et d’indépendance ([liste des organismes](https://www.health-data-hub.fr/sites/default/files/2021-05/202003-%20Engagement_de_conformite_au_reefentiel_au_06-03-20.pdf) ayant déjà fait cet engagement).

## Publication des résultats et valorisation

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Expliquer comment les résultats seront rendus publics : publications scientifiques ou non, communications orales (colloques, revues à comité de lecture, site Internet…).

**S’agissant des données du SNDS (définition en 2.3),** vous devez vous engager à communiquer au grand public vos résultats et la méthodologie associée à l’issue de votre projet en application des obligations de transparence. Ces éléments de résultats seront rendus publics par le biais du [répertoire des projet](https://www.health-data-hub.fr/projets)s sur notre site. N’hésitez pas à consulter [notre guide](https://www.health-data-hub.fr/sites/default/files/2024-05/20240411%20-%20Guide%20p%C3%A9dagogique%20sur%20la%20transparence%20VF%20%281%29.pdf) précisant les modalités de mise en œuvre de cette obligation.

# 

# MÉTHODOLOGIE [3 - 5 PAGES MAXIMUM]

## Design de l’étude

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Décrire le design de l’étude qui sera mise en place (cohorte, cas-témoins…) et indiquer s'il y aura des comparaisons le cas échéant. Préciser s'il s’agit d’analyses transversales ou longitudinales, et si l’étude est prospective ou historique.

Justifier vos choix, leur adéquation avec la question posée/l’hypothèse de recherche.

Décrire les critères qui seront étudiés en précisant quel sera le principal critère et les critères secondaires, le cas échéant.

Si une association entre une exposition et un indicateur de santé est étudiée (par exemple : association entre la consommation de tabac et apparition de maladies respiratoires), indiquer la mesure de l’association qui sera estimée (odds ratio, risque relatif…).

Les avantages du design retenu pour répondre aux objectifs de votre projet peuvent être mentionnés dans ce chapitre.

## Description et justification de la population d’étude

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Décrire la population de l’étude en précisant :

* Les critères d’inclusion et d’exclusion en termes de caractéristiques de la population, de pathologies visées et le périmètre géographique (régional, national, multicentrique etc.) ;
* La période de ciblage ;
* La période d’extraction ;
* Lorsqu'un échantillonnage est effectué à partir d'une population source, il convient de fournir une description de la population source et de décrire la méthode d'échantillonnage. En cas de constitution de groupe témoin, veillez à bien préciser les critères et le ratio qui doivent être appliqués.

Justifier que les données requises sont adéquates et pertinentes au regard des finalités décrites. Les données requises sont limitées à ce qui est nécessaire au regard des finalités décrites en partie 2 - Justification de l’intérêt public : ne pas en demander plus que nécessaire.

**Si votre projet mobilise des données de la base principale du SNDS, l’expression de besoin présente en annexe 2 devra être complétée dans le cadre de la soumission de votre projet.**

|  |
| --- |
| Définition  (RAPPEL : cette partie est à effacer avant envoi du projet) |

**Population d’étude :** population choisie en fonction de ses caractéristiques pour un projet donné

**Cohorte :** population de sujets qui répondent à une définition donnée et qui sont suivis dans le temps

**Extraction :** sous-ensemble de données, à un instant T, issues d’une ou des bases nécessaires à l’étude (registre, cohorte, autres..)

**Période d’extraction :** période sur laquelle seront extraites les données

**Critères de ciblage :** ensembles de règles/filtres permettant de caractériser (“cibler”) la population d’étude dont vous avez besoin pour votre étude.

**Période de ciblage :**  fenêtre temporelle sur laquelle les critères de ciblage sont appliqués pour sélectionner votre population d’étude.

|  |
| --- |
| *Nos conseils*  *(RAPPEL : cette partie est à effacer avant envoi du projet)* |

*Conseil n°1 concernant la méthode de ciblage :*

Préciser autant que possible la méthode de ciblage (algorithme d’inférence des personnes atteintes d’une certaine pathologie en absence d’un code diagnostic par exemple, combinaison de variables etc.).

**Distinguer les données nécessaires pour l’étude elle-même (l’extraction), et celles qui peuvent être nécessaires pour identifier cette population d’étude lorsque différentes (données de ciblage).** En effet, les données de ciblage et l’extraction peuvent différer : par exemple dans un étude visant à étudier des patients hospitalisés pour un AVC de 2012 à 2015 et à analyser leur parcours de soins jusqu’à 2019, la période de ciblage sera 2012 à 2015 et la période d’extraction, 2012 à 2019. Ces données peuvent provenir de différentes sources.

Les critères de ciblage portent aussi bien sur les caractéristiques démographiques de la population (âge, sexe, commune de résidence, etc.) que sur des caractérisitques liées à la consommation de soins ou les hospitalisations (hospitalisation avec code CIM10, délivrance de médicament avec code CIP, etc.).

*Conseil n°2 concernant la justification de la population d’étude :*

Même pseudonymisées, les données de santé restent des données personnelles. Par conséquent, leur accès doit être restreint au strict nécessaire. Une attention particulière sera portée à la complétude de cette partie. Insuffisamment documenté, votre projet pourrait être autorisé mais pour un périmètre plus restreint que celui demandé, ce qui pourrait mettre en risque la fiabilité de l’étude. Soyez donc bien attentif à expliquer pourquoi vous avez besoin de telles données, en particulier quand c’est une donnée sensible au sens du RGPD (par exemple données de santé, données concernant l’origine éthnique), ou des variables particulièrement sensibles au sens de la base principale du SNDS (par exemple code commune de résidence, date de soins complète) ou d’une importante profondeur.

Nous conseillons donc de :

* Justifier bien les périodes requises, pour le ciblage, et pour l’extraction ;
* Fournir le plus de précision possible sur la taille de la population finale : même si vous n’avez pas accès à la base de données, utilisez les données de la littérature ou des données en open source pour faire des estimations et justifier la nécessité de disposer de ce volume ;
* Décrire la représentativité de votre population d’étude ;
* Apporter une attention particulière à justifier l’utilisation de variables particulièrement sensibles (quasi-identifiants).

*Conseil n°3 concernant la définition des critères de ciblage par sous-population :*

Aussi, pour plus de clarté, vous pouvez découper les critères de ciblage par sous-population. S’agissant de la base principale du SNDS, pour remplir l’annexe 2 (relative à l’expression de besoin SNDS) vous pouvez vous appuyer sur le guide pédagogique créé à cet effet et consulter les ressources suivantes pour vous aider : le [dictionnaire des variables](https://health-data-hub.shinyapps.io/dico-snds/), [les données synthétiques au format des données réelles](https://documentation-snds.health-data-hub.fr/formation_snds/donnees_synthetiques.html), [la documentat](https://documentation-snds.health-data-hub.fr/)ion.

En cas de doute sur l'éligibilité de votre étude à une méthodologie de référence, vous pouvez contacter le HDH ou consulter le [site de la CNI](https://www.cnil.fr/fr/traitements-declaration-conformite?field_norme_numerotation_type_value%5B0%5D=6)L.

## Taille de la population

## 

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Fournir une estimation de la taille de la population que vous souhaitez étudier, indiquez les éléments de calculs sous-jacents.

Si vous n’avez aucune idée de la taille de la population envisagée pour l’étude, une estimation théorique est alors attendue soit à partir d’une revue de littérature (y compris des rapports de la HAS, de Santé Publique France…) soit à partir des données disponibles en open data (une section dédiée à ces données est [disponible ici](https://documentation-snds.health-data-hub.fr/open_data/)). A titre d’exemple, le site [ScanSanté](https://www.scansante.fr/) de l’ATIH permet d’estimer le nombre d’hospitalisations à partir des codes diagnostics CIM10 ou les données en open data de l’Assurance Maladie le nombre de bénéficiaires ou le nombre d’exécutions par type de prestation de soins : [Open Damir,](http://open-data-assurance-maladie.ameli.fr/depenses/index.php#Open_DAMIR) [Open Bio](https://www.data.gouv.fr/fr/datasets/open-bio-base-complete-sur-les-depenses-de-biologie-medicale-interregimes/), et [Open Medic](http://open-data-assurance-maladie.ameli.fr/medicaments/index.php) notamment.

## Sources de données

## 

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

L’objectif de cette partie est d’expliquer à quelles données l’équipe projet souhaite accéder, pourquoi et comment.

**Pour la ou les bases de données sources,** il faudra notamment **présenter les éléments suivants** (si pertinent) **:**

* Le nom de la base de données source (exemple : base principale du SNDS ; entrepôt de l’établissement de santé X, registre de l’agence Y, etc.) ;
* Description de la base source et typologies des données disponibles ;
* Données recueillies au sein de cette ou ces bases sources ;
* Profondeur des données à analyser : période d’extraction des données nécessaires à la réalisation de l’étude avant et/ou après la période d’inclusion ;
* Si vous êtes responsable de traitement d’une de ces bases, l’autorisation CNIL associée (exemple : le registre X est autorisé par la CNIL depuis le 19 janvier 2017, autorisation n°1980120, délibération 2017-013…). Si vous n'êtes pas responsable de traitement, le nom du responsable de traitement de la base (excepté pour la base principale du SNDS).

Le cas échéant, justifier le recueil et l’utilisation de données qui pourraient impacter la vie privée en cas de réidentification des patients, à savoir :

* Les données nominatives (nom, prénom) ;
* Le NIR ;
* Les données sensibles au sens du RGPD, telles que des données de santé, ethniques, religieuses, génétiques ou biométriques ;
* Les identifiants potentiels tels que définis par l’article R1461-11 2° du code de la santé publique, à savoir la période de naissance exprimée en mois et année, le code de la commune de résidence et les données infracommunales de localisation, la date des soins, la date du décès et le code de la commune de décès.

**Indiquez si les différentes bases de données sources ont vocation à être appariées ou non.** En cas d’appariement, préciser les variables qui seront pressenties pour le réaliser. La circulation des données pour réaliser l’appariement sera détaillée plus loin dans le protocole scientifique.

## Variables

## 

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Définir les variables de l’étude par type de variables (description de la population, exposition, critères, facteurs de risque, ou autres types de variables y compris les variables de confusion et les potentiels effets modificateurs, le cas échéant).

## Préparation de données

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Indiquer le ou les outils qui seront utilisés pour la préparation de données de l’étude.

Décrire les travaux de préparation des données : nettoyage, mise en qualité, redressement, annotation, dédoublonnement, corrections, etc.

Pour toutes les sources mobilisées dans le cadre de votre projet, vous devrez décrire la méthode de recueil de données, la méthode d’extraction et de préparation de ces données. Les questionnaires ou tout autre mode de recueil peuvent être décrits en annexe.

## Méthodes, traitements et analyses des données

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Indiquer le ou les outils (framework de développement, software, etc) qui seront utilisés pour analyser les données dans le cadre de l’étude ; décrire et justifier ces outils au regard de la question de recherche posée et des données disponibles.

Détailler les méthodes mathématiques ou algorithmiques (statistiques, inférence, modélisation, apprentissage machine, etc.) qui seront utilisées. Par exemple, pour un test statistique décrire le type de test employé (Chi 2, binomial, etc.), pour les analyses multivariées (modèles utilisés, type et critère de choix des variables) ou pour un projet de machine learning la méthode d’apprentissage (SVM, Random forests, réseaux de neurones artificiels récurrents, convolutifs etc.). Bien justifier le choix de la méthode en fonction de la nature des variable considérées, de leur volumétrie et des objectifs du projet.

Pour un projet d’apprentissage machine, décrire comment seront définis les ensembles d’apprentissage, de validation et de test, justifier leur volumétrie, indiquer si des méthodes spécifiques sont utilisées pour l’apprentissage (comme par exemple le bootstrap, le fine-tuning,...).

Si de nouvelles méthodes sont développées dans le cadre du projet ou s’il est prévu d’employer des approches qui sortent des schémas usuels, justifier pourquoi de telles approches sont envisagées et indiquer comment la qualité des résultats pourra être évaluée par comparaison avec l’état de l’art (définir des benchmarks clairs) et si nécessaire d’ajouter les références bibliographiques correspondantes.

Décrire les méthodes de validation et de contrôle des résultats obtenus y compris les analyses de sensibilités, le cas échéant.

## Limites de l’étude

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Identifier les risques et limites de l’étude. Détailler les biais possibles des résultats, ainsi que la méthode prévue pour les contrôler ou les limiter.

Toute limite potentielle du design de l'étude, des sources de données et des méthodes d’analyses, y compris les efforts entrepris pour réduire les biais doit être discutée.

## Calendrier prévisionnel et faisabilité du projet

|  |
| --- |
| Aide au remplissage  (RAPPEL : cette aide au remplissage est à effacer avant envoi du projet) |

Décrire les **grandes étapes du projet et le calendrier** estimé.

**Le calendrier** doit tenir compte :

* Des délais d’obtention des autorisations d’accès de la CNIL et des producteurs de données ;
* Des temps de travaux de préparation des données, de chaînage et d’ingestion des données dans la plateforme ciblée (le cas échéant) ;
* De l’analyse des données et de l’obtention des résultats attendus ;
* Du temps nécessaire à la documentation de la base de données utilisée pour le projet, des programmes et des résultats ;
* Du temps pour valoriser les résultats obtenus, notamment sous forme de publications ou de mise en open source.

**La durée d’accès aux données doit dans ce cadre être précisée :** il s’agit de la durée nécessaire pour réaliser l’étude. Si la durée d’accès aux données est supérieure à celle nécessaire pour réaliser l’étude, préciser la date souhaitée de fin d’accès et justifier votre demande.

**Justifier la faisabilité du projet :** montrer que l’équipe dispose de l’expertise, des moyens humains, des moyens matériels et des ressources pour mener à bien le projet selon le calendrier établi ci-dessus.

Signaler si l’équipe est **formée à l’utilisation des données** visées, de la base principale du SNDS notamment, le cas échéant.

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

Pour étayer le critère de faisabilité, nous vous conseillons de reprendre la liste des membres de l’équipe projet et pour chacun :

* décrire brièvement son apport au projet ;
* son expérience et/ou expertise en la matière (précédent projet, publications phares, formations suivies) ;
* son implication en termes de temps (% ETP).

Les compétences rassemblées dans votre équipe projet doivent couvrir toute la gamme de compétences requises. Pour plus d’information sur l’offre de formation existante relative aux données du SNDS, nous vous invitons à consulter Ia page formation du [site](https://www.health-data-hub.fr/formations).

# PROTECTION DE LA VIE PRIVÉE, SÉCURITÉ ET CONFIDENTIALITÉ DES DONNÉES [10 PAGES MAXIMUM] (Section destinée à la CNIL)

## Respect des droits des personnes concernées

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Dans le cadre d’un projet utilisant des données à [caractère personnel](https://www.cnil.fr/fr/cnil-direct/question/une-donnee-caractere-personnel-cest-quoi), les principes relatifs à la protection des données s’appliquent. Pour permettre aux personnes concernées par les données du projet d’en garder la maîtrise, ces dernières ont plusieurs droits (se reporter à la partie ci-dessous Définitions pour connaître tous les [droits des personnes](https://www.health-data-hub.fr/sites/default/files/2021-01/Mes_droits.pdf) vis-à-vis des données).

La difficulté pour renseigner cette partie provient en premier lieu de la nécessité d’identifier la situation dans laquelle vous vous trouvez (conformez-vous à la partie nos conseils).

|  |
| --- |
| ***Définition***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

Les droits des personnes vis-à-vis des données :

**L’information** aux personnes concernées doit être exhaustivement conforme au [Règlement Général sur la Protection des Données](https://www.cnil.fr/fr/reglement-europeen-protection-donnees) (RGPD) et à la [Loi Informatique et Libertés](https://www.cnil.fr/fr/la-loi-informatique-et-libertes) (LIL). Elle doit être individuelle, préalable à la mise en œuvre de la recherche et spécifique à chaque projet. L’obligation d’information revient au responsable de traitement.

Le **droit d‘accès** permet à toute personne concernée par les données de demander à consulter les données la concernant et à en obtenir une copie.

Le **droit de rectification** permet à toute personne concernée par les données de demander de les faire corriger si elle constate qu’elles contiennent une erreur.

Le **droit à l’effacement** permet à toute personne concernée par les données de demander à ce qu’elles soient effacées. Ce droit n’est pas forcément applicable, notamment lorsqu’il compromet gravement la réalisation de la recherche, mais cela doit être mentionné dans la note d’information relative à l’étude.

Le **droit d’opposition** permet à toute personne ne souhaitant pas que les données la concernant servent à l’étude, de s'opposer à cette utilisation. Cette opposition empêche toute utilisation ou conservation de données concernée

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Je suis / Je mène** | ***Capacité à contacter les personnes concernées*** | **Je mène une étude rétrospective** | **Je mène une étude prospective** | **Je mène les deux** |
| **Je suis en charge de la base de données que j’utilise, c’est moi qui vais procéder à l’information si je suis en capacité de contacter les personnes concernées** | Oui | Je fournis les caractéristiques de l’information individuelle à laquelle j’ai/je vais procédé/er (conditions, vecteurs, lettres d’information, dates) | Je fais une information individuelle | Je fais une information individuelle |
| Non | Je demande une dérogation (j’argumente) | Je ne peux pas réaliser l’étude | Je ne peux pas faire l’étude |
| Oui et Non | NA | NA | Je peux contacter les personnes concernées pour l’étude prospective mais pas pour l’étude rétrospective : je fais l’information des premiers et demande une dérogation pour les seconds |
| **Je ne suis pas en charge de la base de données que j’utilise, je contacte le responsable de la base (RD) et je détermine l’action à réaliser en fonction de sa capacité à contacter les personnes concernées** | Oui | Le RD me fournit les caractéristiques de l’information individuelle à laquelle il a procédé/va procéder (conditions, vecteurs, lettres d’information, dates) | Le RD me fournit les caractéristiques de l’information individuelle à laquelle il va procéder (conditions, vecteurs, lettres d’information, dates) | Le RD me fournit les caractéristiques de l’information individuelle à laquelle il a/va procédé/er (conditions, vecteurs, lettres d’information) |
| Non | Je demande une dérogation | Je ne peux pas réaliser mon étude | Je ne peux pas réaliser l’étude |
| Oui et Non | NA | NA | Le RD contacte les personnes concernées pour l’étude prospective mais pas pour l’étude rétrospective : il me fournit les informations relatives à l’information des premiers et je demande une dérogation pour les seconds |

**Remarques :**

* Pour rappel, lorsque je ne suis pas en charge de la base de données que j’utilise, l’obligation d’information repose quand même juridiquement sur moi : je dois faire le nécessaire pour que les personnes soient informées ;
* Si l’acteur en charge des données ne me répond pas ou si j’ai des questions, je peux me tourner vers le Health Data Hub pour voir si je peux débloquer la situation ([dpd@health-data-hub.fr](mailto:dpd@health-data-hub.fr)).

**Quelles informations devez-vous fournir sur l’information que vous avez faite ou que le RD a faite ou va faire :**

* Indiquer les mesures qui ont été ou seront mises en place pour assurer l’information individuelle des personnes concernées par les données utilisées pour le projet (les délais, les moyens - email, voie postale…)
* Détailler le contenu de la lettre d’information individuelle initialement reçue par les personnes concernées (et le cas échéant, joindre la lettre) et indiquer la période durant laquelle cette information a été délivrée.

Le contenu d’une lettre d’information est le suivant :

* L’identité et les coordonnées du responsable de traitement (nom, prénom, structure, email, adresse postale…) ;
* La finalité du traitement de données (présentation synthétique et claire du projet de recherche) ;
* La base juridique du traitement au sens de l’article 6 du RGPD (par exemple obligation légale, exécution d’une mission d’intérêt public, intérêts légitimes du responsable de traitement) ;
* La base juridique du traitement des données sensibles (par exemple les données de santé, les données concernant l’origine ethnique) au sens de l’article 9 du RGPD (par exemple recherche scientifique) ;
* La durée de conservation des données à caractère personnel ou les critères utilisés pour déterminer cette durée. La durée indiquée doit être déterminée ou déterminable (par exemple 2 ans pour l’analyse des données, puis 3 ans pour la publication) ;
* Les destinataires ou les catégories de destinataires des données (les personnes qui auront accès aux données dans le cadre de la recherche, y compris les sous-traitants) ;
* Le droit de demander l'accès aux données, la rectification ou l'effacement de celles-ci ou la limitation du traitement relatif à la personne concernée et le droit de s’opposer à l’utilisation des données ;
* Les modalités d’exercice de ces droits ;
* Le cas échéant, le transfert de données à caractère personnel hors de l’Union européenne et la référence aux garanties appropriées et les moyens d’en obtenir une copie ou l’endroit où elles ont été mises à disposition ;
* Le cas échéant, [l’existence d’une prise de décision automatisée, y compris un profilage ;](https://www.cnil.fr/fr/profilage-et-decision-entierement-automatisee)
* Le droit d’[introduire une réclamation auprès de la Commission Nationale de l’Informatique et des Libertés (CNIL)](https://www.cnil.fr/fr/agir) ;
* Les coordonnées (email et postale) du délégué à la protection des données (DPD) du responsable de traitement ;

Pour les études rétrospectives, deux mentions d’information supplémentaires doivent figurer dans la lettre :

* La source d'où proviennent les données à caractère personnel (la base de données utilisée) ;
* La nature des informations qui seront utilisées dans la recherche (les types de données comme les données médico-administratives, les données de santé, les données d’habitude de vie...).

Pour les études prospectives, vous devez :

* Indiquer si les personnes sont obligées de fournir leurs données personnelles ou non (ex. astérisques identifiant les champs obligatoires dans un questionnaire) et les conséquences si elles ne les fournissent pas (ex. leur préciser que dans tous les cela n’impactera pas leur prise en charge) ;
* Préciser le caractère facultatif de la participation de la personne à cette étude (les personnes sont informées qu’il n’y a pas d’obligation de participer, qu’elles peuvent s’opposer à ce que leurs données soient utilisées sans conséquences particulières).

**En cas de demande de dérogation à l'information :**

* Fournissez et détaillez le motif choisi entre :
  + la fourniture des informations se révèle impossible (par exemple impossibilité de retrouver l’identité des personnes, nombre de personnes trop important, etc.) ;
  + la fourniture des informations exigerait des efforts disproportionnés (attention la CNIL estime que le coût ne constitue pas une justification suffisante), ou
  + l’obligation d’information est susceptible de rendre impossible ou de compromettre gravement la réalisation des objectifs de la recherche ;
* Détaillez les efforts que cela nécessiterait pour communiquer les informations à la personne concernée. Ces efforts doivent être comparés à l’incidence et aux effets que l’absence d’information pourrait avoir sur la personne concernée ;
* Détailler les efforts mis en oeuvre pour contacter le responsable de données le cas échéant ;
* Détailler les garanties que vous mettez en place pour compenser l’absence d’information individuelle et remplir votre obligation d’information collective conformément à l’article 14-5b) du RGPD (publication d’une information générale sur le site web du responsable de traitement, des responsables de données, du Health Data Hub, affichage dans les locaux du responsable de données, mention dans des supports spécifiques, dans un livret d’accueil, dans une newsletter, communication par voie de presse, communication aux associations de patients) ;
* Intégrer en annexe les documents relatifs à l'information collective que vous prévoyez de réaliser.

Pour vous aider dans la formalisation de cette information (qu’elle soit individuelle ou collective), un modèle de note d’information est mise à votre disposition dans le [starterkit](https://www.health-data-hub.fr/actualites/kit-pedagogique).

**A noter :** Il est possible de demander une dérogation partielle à l’obligation d’information, par exemple pour les patients dont le suivi est achevé ou une population trop ancienne. Cependant, il conviendra de détailler les modalités d’information individuelle pour la partie de la population qu’il est possible d’informer.

**Information lorsque vous mobilisez uniquement la base principale du SNDS :** Lorsqu’une étude porte uniquement sur les données de la base principale du Système national des données de santé (SNDS), il convient de demander une dérogation à l’obligation d’information individuelle sur la base de l’article 14.5 b) du RGPD, au motif que l'information se révèle impossible. Par exemple :

*"Conformément au code de la santé publique, les données de la base principale du SNDS ont fait l’objet de mesures de pseudonymisation par un procédé cryptographique irréversible, de sorte qu'il est impossible d'identifier directement les personnes concernées. Ainsi, dans le cadre de l’étude, l'information individuelle des patients se révèle impossible et une dérogation à cette obligation est demandée conformément aux dispositions de l'article* 14.5 b) *du RGPD."*.

Cette dérogation doit s’accompagner d’une information collective propre à l’étude. La CNIL recommande que la note d’information associée soit a minima publiée sur le site web du responsable de traitement. Il est également obligatoire d’inscrire le projet sur le répertoire public du HDH.

De plus, lorsque le responsable de traitement réalise plusieurs études à partir des données de la base principale du SNDS, la CNIL recommande dans ses MR007 et MR008 (relatives aux projets utilisant exclusivement les données de la base principale du SNDS) la mise en place d'un portail de transparence comportant une information générale sur la base principale du SNDS, puis une note d'information spécifique à chaque projet mobilisant les données de celle-ci.

**Autres droits : L’exercice des droits des personnes concernées (accès, rectification, limitation, effacement, opposition)**

Les procédures décrites doivent êtres celles de votre projet spécifique, objet du présent protocole.

La situation diffère selon que vous êtes en charge de la base ou non :

* Si vous n’êtes pas à l’origine de la collecte, la personne concernée ne peut très probablement pas exercer ses droits directement auprès de vous. Néanmoins, nous vous conseillons de mettre en visibilité un contact de DPD (délégué à la protection des données, appelé aussi DPO) sur votre fiche projet dans le répertoire public des projets pour pouvoir leur indiquer la procédure à suivre s’ils vous contactent en direct. Vous devrez vous renseigner auprès du responsable de la base pour connaître la procédure en place ou les dérogations valables ;
* Si vous êtes à l’origine de la collecte, décrivez les procédures en place.

Dès lors que la base principale du SNDS est utilisée dans le cadre d’une étude (seule ou appariée à d’autres données), l’exercice des droits s’effectue auprès du responsable de traitement de cette étude. Pour l’exercice des droits sur la base principale du SNDS en tant que telle (en dehors de tout projet), la demande doit être adressée au directeur de l'organisme gestionnaire du régime d'assurance maladie obligatoire auquel la personne est rattachée, conformément à l’article R1461-9 du code de la santé publique.

Pour les utilisateurs de la plateforme technologique du Health Data Hub : Les coordonnées des DPD des responsables de données des bases disponibles sont disponibles sur le site internet du Health Data Hub. Si le Health Data Hub est contacté directement par une personne, il l’invite à contacter le(s) producteur(s) de données qui est/sont seul(s) à disposer des identifiants directs des personnes et qui peut / peuvent donc donner suite à un exercice des droits.

## Confidentialité et sécurité des données

### Gestion du risque de réidentification

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

Dans cette partie, il s’agit de préciser comment est mitigé le risque de ré-identification des personnes concernées (les termes marqués d’un astérisque sont définis dans la section “Définition”) :

* Les données sont-elles pseudonymisées\* (et si oui, comment), ou des données directement identifiantes\* (noms, prénoms) sont-elles utilisées ?
* Les données indirectement identifiantes\* (NIR, adresse de résidence, dates exactes de soin de soin, de naissance ou de décès, données génétiques) sont-elles généralisées\*, et si oui, comment ?
* Des données ethniques, religieuses, philosophiques ou syndicales, génétiques, données biométriques, sur la vie ou l’orientation sexuelle sont-elles utilisées ?
* Combien de temps les données sont-elles conservées ? dit autrement, une fois l’objectif de traitement des données atteint, combien de temps sont-elles conservées par l’hébergeur de données (justifiez bien ce point). Vous trouverez [ici](https://www.cnil.fr/fr/les-durees-de-conservation-des-donnees), les recommandations de la CNIL sur la durée de conservation des données. La durée de conservation englobe la durée d’accès aux données.

|  |
| --- |
| ***Définition***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Pseudonymisation (masquage) :** Remplacement irréversible d’une donnée ou un ensemble de données par un numéro non signifiant (e.g. remplacement du NIR par un numéro de cohorte aléatoire).

**Identifiant direct :** Un identifiant direct est une donnée qui peut être directement liée à un individu unique (e.g. nom, prénom).

**Identifiant indirect :** Un identifiant indirect, est une donnée qui, prise isolément, ne permet pas de réidentifier uniquement une personne, mais dont la combinaison avec d’autres données (notamment les variables sensibles) le permet (e.g. NIR, code postal de résidence, date de naissance, dates d'événements précis).

**Généralisation :** Réduction de la granularité (finesse, niveau de détail) d’une variable (e.g. remplacement du code commune par le numéro du département).

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

Il est possible que vous ne soyez pas directement en mesure de répondre à ces questions, faute d’avoir accès vous-mêmes aux données. Dans ce cas, nous vous recommandons de vous rapprocher du Responsable de Données qui sera en mesure de vous indiquer les mesures mises en place pour réduire le risque de réidentification.

Pour les cas précis de la base principale du SNDS et des bases intégrées au catalogue du Health Data Hub, vous pouvez vous inspirer des réponses suivantes, qui sont à adapter spécifiquement à votre projet.

**Pour la base principale du SNDS :**

* Les données de la base principale du SNDS sont pseudonymisées et ne contiennent aucune donnée directement identifiante. L’identifiant patient utilisé est dérivé du NIR de manière irréversible par applications successives des différentes fonctions de pseudonymisation ;
* Parmi les variables quasi-identifiantes, les suivantes sont généralisées :
* La date de naissance est tronquée pour ne conserver que l’année et le mois ;
* L’adresse de résidence est généralisée au code commune.
* La base principale du SNDS ne contient aucune donnée liée à l’affiliation ethnique, politique, religieuse, philosophique ou syndicale, ni donnée génétique, ni donnée biométrique, ni donnée sur la vie ou l’orientation sexuelle.

### Support de données

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

L’objectif de cette partie est de présenter les différents supports sur lesquels les données vont être traitées ainsi que les principales mesures techniques et organisationnelles visant à assurer leur sécurité.

Pour les demandes d’accès au SNDS, vous devez respecter le [référentiel de sécurité du SNDS](https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000049516244). Si vous menez votre étude sur le portail de la Cnam, vous n’avez pas d’information particulière à fournir en ce sens.

Dans le cas où votre étude se déroule sur la [plateforme technologique du Health Data Hub,](https://www.health-data-hub.fr/page/notre-offre-technologique) cette dernière a fait l’objet d’une homologation.Vous pouvez reprendre dans cet emplacement du protocole le paragraphe suivant : “L’analyse des données aura lieu sur un espace projet dédié, hébergé sur la plateforme technologique du Health Data Hub, qui est conforme à la Politique de sécurité des systèmes d'information de l’Etat et plus spécifiquement [le référentiel de sécurité](https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034265125&amp;categorieLien=id) du SNDS. Le contexte de sécurité de la plateforme technologique du Health Data Hub a été défini dans le cadre d’une démarche conduite en collaboration avec l’Agence nationale de la sécurité des systèmes d’information (ANSSI) et le Haut fonctionnaire de défense et de sécurité du Ministère des Solidarités et de la Santé. Dans ce contexte, des mesures de sécurité aussi bien organisationnelles que techniques sont mises en œuvre par le HDH pour protéger l’accès aux données de santé.”

En revanche, si vous utilisez une autre plateforme d’accès (système fils /bulle sécurisée homologuée), vous devez expliquer ici en quoi ce système est conforme au référentiel [de sécurité du SNDS](https://www.snds.gouv.fr/download/Guide_accompagnement.pdf) (homologation comprise).

### Circuit des données en cas d’appariement

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

L’objectif de cette partie est de démontrer que la circulation des données envisagée est cadrée, sécurisée, et qu’elle préserve la confidentialité des données.

Il s’agit donc de décrire la circulation des données des bases de données sources à la cohorte finale, en détaillant les flux de données, les acteurs impliqués, les durées de conservation des données et des tables de correspondance le cas échéant.

Il faut indiquer qui aura accès aux différentes sources de données, aux différentes étapes de la procédure.

Pour plus d'informations sur les processus d’appariement, vous pouvez vous référer aux [schémas d’appariement et de circulation des données](https://www.health-data-hub.fr/sites/default/files/2021-05/Sch%C3%A9mas%20d%E2%80%99appariement%20et%20de%20circulation%20des%20donn%C3%A9es_HDHxCNAM.pdf) et les conseils ci-dessous ou consulter, en cas de recours au NIR, [le guide pratique](https://www.cnil.fr/fr/les-circuits-dappariement-avec-le-snds-utilisant-le-nir) produit par la CNIL.

Pour le responsable de traitement ne pouvant disposer lui-même des données directement identifiantes nécessaires à un appariement (NIR, nom et prénom par exemple), mais que ces données sont détenues par une autre partie, il sera nécessaire de recourir à un intermédiaire indépendant (tiers de confiance) qui sera le garant de la confidentialité des identifiants utilisés.

|  |
| --- |
| ***Définition***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

**Appariement direct par l’utilisation du NIR** : cela consiste à rapprocher les données issues de différentes sources à partir d’un élément identifiant (NIR) partagé par ces sources. Si vous avez la possibilité de recourir au NIR, cet appariement est à privilégier sur celui consistant à le reconstruire.

**Appariement direct avec reconstruction du NIR** : cela consiste, à partir de certaines variables identifiantes présentes dans une source à reconstituer le NIR via la mobilisation d’un tiers habilité. Cet identifiant pourra ensuite être rapproché du NIR présent dans l’autre source.

**Appariement indirect** : cela consiste à apparier les données issues de différentes sources à partir de données communes (non directement identifiantes) à ces sources.

|  |
| --- |
| ***Nos conseils***  *(*RAPPEL : cette partie est à effacer avant envoi du projet*)* |

Il existe un grand nombre de possibilités mais nous avons rassemblé ci-dessous les schémas de circulation des données les plus communs.

**Cas n°1** - **Appariement indirect d’une cohorte avec la base principale du SNDS**

**Etape 1 - Constitution de la cohorte initiale et de la table des variables communes pour l’appariement**

A partir de la base source, le producteur constitue la cohorte initiale [A], comme décrit au paragraphe 3.1. « Données requises ». Un numéro de cohorte non signifiant (ID\_temp) est attribué de façon aléatoire à chaque personne concernée.

Pour préparer l’appariement indirect des données avec la base principale du SNDS, le producteur crée une table des variables communes intégrant un numéro d’appariement non signifiant (ID\_temp) - créé spécifiquement et différent de l’identifiant communément présent dans les données source - et les variables communes entre la cohorte A et la base principale du SNDS, telles que :

* Le code FINESS de l’hôpital ;
* La date d’entrée et de sortie ;
* Le mois et l’année de naissance ;
* Le sexe ;
* Le diagnostic principal (DP), le diagnostic relié (DR) et les diagnostics associés significatifs (DAS) ;Le mode d’entrée et de sortie ;
* Les codes CCAM, si existants.

Cette table des variables communes sera déposée sur une plateforme sécurisée de la Cnam.

**Etape 2 - Extraction des données de la base principale du SNDS**

A partir des variables contenues dans la table des variables communes, la Cnam effectue une recherche des personnes concernées au sein de la base principale du SNDS et développe l’algorithme d’appariement pour répondre au besoin de l’étude. Un flowchart est réalisé par la Cnam et est validé avec le demandeur. La Cnam réalise ensuite l’extraction des données nécessaires pour le projet. Par exemple, elle va rechercher l’ensemble des données relatives aux consommations de soins ambulatoires et hospitalières de la population ciblée sur les 3 dernières années. Elle attribue un identifiant non signifiant spécifique au projet (ID\_SNDS\_proj) et consolide une table de correspondance des identifiants pseudonymisés [ID\_temp ; ID\_SNDS\_proj] entre l’identifiant temporaire transmis par le producteur (ID\_temp) et l’identifiant présent dans l’extraction de la base principale du SNDS qui sera fournie (ID\_SNDS\_proj).

La Cnam conserve les correspondances entre les pseudonymes le temps des échanges avec le demandeur, visant à valider l’appariement. Une fois l’extraction de la base principale du SNDS et la table de correspondance des identifiants pseudonymisés [ID\_temp ; ID\_SNDS\_proj] transmises au Health Data Hub, à l’espace projet dans le portail de la Cnam, ou à la bulle sécurisée du RT ou du prestataire choisi, la Cnam supprime les correspondances entre les pseudonymes et la table des variables communes (sauf si une mise à jour est prévue auquel cas la table est conservée jusqu’à la fin de la durée de l’étude).

**Etape 3 – Appariement et traitement des jeux de données dans la plateforme de réception des données**

Si vous accédez aux données dans un espace projet du portail de la Cnam, ces dernières seront accessibles sur l’espace projet dédié.

Si vous accédez aux données dans un système fils/bulle sécurisée (autre que la plateforme du HDH), le producteur transmet les données suivant les modalités convenues. La Cnam transmettra l’extraction de la base principale du SNDS via une modalité choisie par le gestionnaire parmi les propositions de la Cnam. Si vous accédez aux données sur la plateforme du HDH, la cohorte A pseudonymisée et l’extraction de la base principale du SNDS sont transmises respectivement par le producteur et la Cnam au gestionnaire de la plateforme cible via des réseaux dédiés, permettant un transfert privé (hors internet) et sécurisé des données.

Dans tous les cas, les données sont stockées et traitées dans un espace projet sécurisé, accessible aux personnes habilitées seulement. Elles seront conservées pendant le nombre d’années autorisées pour la réalisation de l’étude et la publication des résultats, puis elles seront archivées pendant un nombre d'années défini par l’autorisation puis seront supprimées définitivement.

**Cas n°2** - **Appariement direct avec la base principale du SNDS par l’utilisation du NIR pour un traitement prévu [dans une plateforme cible]**

**Etape 1 - Constitution de la cohorte initiale et transmission du NIR à la Cnam**

Dans le cadre de sa base source, le producteur dispose pour chaque personne de son NIR en clair, de sa date de naissance et de son sexe.

A partir de la base source, le producteur constitue la cohorte initiale [A], comme décrit au paragraphe 3.1.

« Données requises ». Un numéro de cohorte non signifiant (ID\_temp) est attribué de façon aléatoire à chaque personne concernée.

Le producteur transmet à une structure déconcentrée de la Cnam la table de pseudonymisation [ID\_temp ; NIR] constituée du NIR du patient et/ou de l’ouvrant droit, du sexe, de la date de naissance ainsi qu’un numéro de cohorte non signifiant (ID\_temp) pour toutes les personnes présentes dans la cohorte A créé spécifiquement et différent de l’identifiant communément présent dans les données source. Le producteur devra créer un fichier normé à partir de la table de pseudonymisation [ID\_temp, NIR] et le déposer sur la plateforme SAFE suivant les modalités qui lui seront communiquées ultérieurement

**Etape 2 - Extraction des données de la base principale du SNDS à partir du NIR**

Une structure déconcentrée de la Cnam applique aux NIR en clair figurant dans le fichier normé (généré à partir de la table de pseudonymisation [ID\_temp ; NIR]) les mêmes algorithmes de hachage utilisés lors de la consolidation de la base principale du SNDS. A partir de l’identifiant SNDS ainsi obtenu, la Cnam cible les personnes concernées de l’étude dans la base principale du SNDS et extrait les données requises pour le projet. Elle attribue un identifiant non signifiant spécifique au projet (ID\_SNDS\_proj) et consolide en parallèle une table de correspondance des identifiants pseudonymisés [ID\_temp ; ID\_SNDS\_proj] entre l’identifiant temporaire transmis par le producteur (ID\_temp) et l’identifiant présent dans l’extraction de la base principale du SNDS qui sera fournie (ID\_SNDS\_proj).

La Cnam transmet ensuite cette table de correspondance des identifiants pseudonymisés [ID\_temp ; ID\_SNDS\_proj] ainsi que l’extraction de la base principale du SNDS au gestionnaire de la plateforme cible. Les données extraites sont supprimées à l’issue du processus mais la correspondance entre les pseudonymes pourra être conservée par la Cnam pendant la durée d’accès du projet aux données de la base principale du SNDS, notamment si des mises à jour des données de cette base sont nécessaires.

**Etape 3 – Appariement des jeux de données et traitement dans la plateforme cible**

La cohorte A pseudonymisée et l’extraction de la base principale du SNDS avec la table de correspondance des identifiants pseudonymisés [ID\_temp ; ID\_SNDS\_proj] sont transmises respectivement par le producteur et la Cnam au gestionnaire de la plateforme cible via des réseaux dédiés, permettant un transfert privé (hors internet) et sécurisé des données.

A partir du numéro de cohorte non signifiant (ID\_temp), les jeux de données sont ensuite appariés dans cette plateforme par l’opérateur de la plateforme et un nouvel identifiant est créé de façon aléatoire pour chaque personne concernée (ID\_projet).

Les données sont stockées et traitées dans un espace projet sécurisé de la plateforme cible, accessible aux personnes habilitées seulement. Elles seront conservées pendant le nombre d’années autorisées pour la réalisation de l’étude et la publication des résultats, puis elles seront définitivement supprimées.

# ANNEXES

|  |
| --- |
| ***Aide au remplissage***  *(*RAPPEL : cette aide au remplissage est à effacer avant envoi du projet*)* |

**Annexes :**

* Notice d’information soumise aux personnes concernées le cas échéant ;
* Liste des variables utilisées et l’expression de besoin de la base principale du SNDS (si pertinent) ;
* Bibliographie.

**Annexe 1 - Exemple fictif de protocole**

(A supprimer avant envoi au CESREES)

**L’encéphalopathie de type 1 :**

**une étude clinique appariée aux données du Système Nationale des Données de Santé (SNDS)**

Sommaire de l’exemple du protocole scientifique

**PRÉSENTATION DE L’ÉQUIPE PROJET [1 PAGE MAXIMUM] 29**

**OBJECTIFS ET FINALITÉS [2 - 4 PAGES MAXIMUM] 30**

Contexte, objectif(s) et justification de l’étude 30

Justification du respect de l’éthique 31

Justification de l’intérêt public 31

Publication des résultats et valorisation 31

**MÉTHODOLOGIE [3 - 5 PAGES MAXIMUM] 32**

Design de l’étude 32

Description et justification de la population d’étude 32

Taille de la population 33

Sources de données 33

Variables 34

Préparation de données 34

Méthodes, traitements et analyses des données 35

Limites de l’étude 35

Calendrier prévisionnel et faisabilité du projet 36

**PROTECTION DE LA VIE PRIVÉE, SÉCURITÉ ET CONFIDENTIALITÉ DES DONNÉES [10 PAGES MAXIMUM] (Section destinée à la CNIL) 37**

Respect des droits des personnes concernées 37

Confidentialité et sécurité des données 37

Gestion du risque de réidentification 37

Support de données 38

Circuit des données en cas d’appariement 38

**ANNEXES 28**

1. **Présentation de l’équipe projet [1 page maximum]**

**RESPONSABLE DE TRAITEMENT**

Le responsable de traitement est le laboratoire pharmaceutique PharmaX basé à Paris en France. PharmaX est représenté par les personnes suivantes :

X, Président de PharmaX.

* Son rôle sera de s’assurer du bon déroulement de l’étude dans les délais alloués au projet. Il sera le signataire de la convention de collaboration avec le responsable de mise en œuvre et le health data hub.

X, Chef de Projet Clinique.

* Son rôle sera de mettre en place, de suivre et de coordonner le projet ;

Le responsable de traitement PharmaX est aussi le financeur pour ce projet.

**DÉLÉGUÉ À LA PROTECTION DES DONNÉES**

X, Délégué à la protection des données, est également Directrice des Affaires Réglementaires et Responsable de l’Assurance Qualité, au sein de l’organisation PharmaX

* Son rôle sera de s'assurer que les données soient traitées dans le respect de la confidentialité et conformément à la RGPD ;

**RESPONSABLE DE LA MISE EN ŒUVRE DU TRAITEMENT**

Le responsable de la mise en œuvre est le bureau d’étude HealthX situé à Paris. Il est représenté par XX, directeur général. HealthX a déjà conduit plusieurs projets utilisant les données de la base principale du Système National des Données de Santé historique (SNDS historique). Pour ce projet en particulier, les membres de l’équipe projet sont les suivants :

* X, Directeur Etude Vie Réelle et Co-dirigeant de HealthX, Bureaux d’études partenaire de qui sera le Responsable de la mise en œuvre du projet.
* X, biostatisticienne et spécialiste des données de la base principale du SNDS

**COMITÉ SCIENTIFIQUE**

Un comité scientifique indépendant a également été mis en place dans le cadre de ce projet. Son rôle est de revoir et valider le protocole de l’étude, et de participer à l’interprétation et la discussion des résultats. Ce comité scientifique est composé des personnes suivantes :

* Pr X, MD, biostatisticien et pharmaco-épidémiologiste,
* Pr X, MD, neurologue,
* Pr X, économiste de la santé
* Mme X, représentante d’une association d’usager du système de santé

1. **Objectifs et finalités [2 - 4 pages maximum]**
   1. **Contexte, objectif(s) et justification de l’étude**

**Contexte**

L’encéphalopathie de type 1 (EGT1), dans sa forme sévère, est une encéphalopathie caractérisée par une épilepsie de l'enfant résistante au traitement, une décélération de la croissance du périmètre crânien résultant en une microcéphalie, un retard du développement psychomoteur, une spasticité, une ataxie, une dysarthrie et d'autres troubles neurologiques paroxystiques survenant souvent avant les repas (référence bibliographique).

Les premiers symptômes peuvent apparaître dès 3 mois, après une naissance et une gestation normale. Dans ses formes mineures, modérées et/ou anciennes, dites atypiques, un retard psychomoteur associé ou non à un handicap est fréquemment observé (référence bibliographique).

Maladie pédiatrique handicapante, elle bénéficie pourtant d’un traitement qui repose sur un régime alimentaire adapté, à savoir un régime cétogène. D’autres approches thérapeutiques innovantes sont par ailleurs en cours d’étude (référence bibliographique).

Le besoin majeur pour cette maladie est actuellement de poser un diagnostic correct et le plus précocement possible pour faire bénéficier les patients des options thérapeutiques. On estime à 800 le nombre de patients prévalents en France (référence bibliographique), dont seulement un peu plus d’une centaine seraient diagnostiqués (référence bibliographique).

Aujourd’hui, le diagnostic repose en premier lieu sur le dosage de la glycorachie dans le liquide céphalorachidien couplé au dosage de la glycémie. Cette première étape nécessite la réalisation d’une ponction lombaire, nécessitant une hospitalisation et n’est pas systématiquement réalisée du fait de son caractère invasif et de la complexité de sa réalisation.

Une analyse génétique complémentaire peut être demandée. Cette analyse moléculaire est complexe à réaliser du fait du polymorphisme génétique de cette affection. Elle permet de poser un diagnostic de certitude pour les mutations connues. En revanche un nombre significatif de mutations non contributives ne permet pas de conclure à un EGT1 avéré. L’interprétation du généticien n’est disponible en moyenne qu’au bout de 3 mois.

**Justification et finalité de l’étude**

Le diagnostic de cette pathologie repose donc aujourd’hui sur un faisceau d’arguments. La complexité de la démarche diagnostique est une des raisons majeures du sous-diagnostic de cette maladie (100 cas diagnostiqués sur 800 sujets affectés estimés en France) et d’une importante errance thérapeutique (référence bibliographique).

Le laboratoire PharmaX a développé un nouveau test d’aide au diagnostic de l’affection considérée, le test Alfa. Ce test est beaucoup plus simple à réaliser que la ponction lombaire. Il s’effectue sur sang total (0,5 mL est suffisant) et le résultat peut être disponible en 48 heures.

Les travaux effectués ont validé l’excellente répétabilité et reproductibilité du test Alfa. Les performances analytiques obtenues sur une cohorte rétrospective de 30 patients ont montré une sensibilité voisine de 78 % et une spécificité > à 99,5 (référence bibliographique). Ses performances analytiques ont conduit au marquage CE en janvier 2013.

Le test Alfa avait fait l’objet d’une étude clinique Beta (décrite dans la partie méthode), visant à confirmer ses performances sur une large échelle.

L’étude visée par ce protocole est une étude longitudinale rétrospective qui sera réalisée d’une part à partir des données de la base principale du SNDS, et d’autre part à partir d’un appariement indirect des données cliniques qui ont été collectées dans l’étude Beta avec les données de la base principale du SNDS. Cette étude aura pour finalité de mieux connaître les caractéristiques des patients atteints de l’EGT1 pour améliorer et faciliter le diagnostic de cette pathologie et d’en améliorer sa prise en charge.

Le projet sera supervisé par un comité scientifique composé d’un biostatisticien et pharmaco-épidémiologiste, un neurologue, un économiste de la santé et une représentante d’une association d’usager du système de santé.

**Objectifs de l’étude**

1 - L’objectif principal de l’étude sera d’une part d’étudier la prévalence de l’EGT1 sur les années 2010-2015 à partir des données de la base principale du SNDS.

2 - Les objectifs secondaires de l’étude sont :

* 2.1 - Estimer, de façon annuelle, les ressources consommées et les coûts des soins de santé des patients atteints de l’EGT1 jusqu’en 2019 ;
* 2.2 - Estimer le délai entre la date des premières manifestations patentes de troubles neurologiques et la date du diagnostic de l’EGT1 (errance médicale).

La prévalence de l’EGT1 et la mortalité toutes causes confondues seront estimées à partir des données de la base principale du SNDS. Les objectifs 2.1 et 2.2 seront analysés à partir des données appariées (appariement indirect). Deux cohortes seront donc mises en place, elles seront décrites dans la partie méthodologie.

Les critères d’évaluation des objectifs faisant l’objet du présent protocole reposeront essentiellement sur une description de paramètres quantitatifs (prévalence, mortalité toutes causes confondues, coûts, délais…) en comparant les résultats obtenus avec les données de la littérature.

* 1. **Justification du respect de l’éthique**

Le projet ne pose pas de problème éthique, pas de stigmatisation d’un groupe spécifique et ne va pas à l’encontre de la morale. Cette étude est réalisée dans le contexte d’une amélioration de la prise en charge des patients de la maladie rare (prévalence estimée à 1 cas pour 80 000 en France) EGT1. L’étude clinique inclut des patients sur la base du volontariat, après que ces patients aient reçu une information complète et adéquate et après qu’ils aient donné leur consentement écrit. De plus, les patients ont également donné leur consentement pour l’utilisation de leurs données du SNDS pour effectuer l’appariement.

* 1. **Justification de l’intérêt public**

Le test Alfa répond à un besoin médical insuffisamment couvert actuellement. La mobilisation des données de l’étude clinique chaînée aux données de la base principale du SNDS va permettre d’acquérir beaucoup de connaissances supplémentaires sur cette pathologie, le diagnostic des patients EGT1 pourra en être amélioré dans le bilan étiologique des troubles neurologiques inexpliqués en pédiatrie et chez le jeune adulte.

Cela doit avoir comme conséquence une réduction du délai de diagnostic de certitude des patients EGT1.

La diminution des coûts de la prise en charge pour une efficacité clinique équivalente (ou supérieure) est également un bénéfice important.

En résumé, la connaissance de l’épidémiologie de la maladie, du parcours de soins des patients, des ressources de santé utilisées ainsi que leurs coûts associés permettra de mieux caractériser les besoins en termes de diagnostic et donc d’améliorer la prise en charge des patients.

* 1. **Publication des résultats et valorisation**

Conformément aux Recommandations de Déontologie et Bonnes Pratiques en Epidémiologie et à la Déclaration d’Helsinki, les résultats de l’étude seront soumis à publication. Au moins un article scientifique sera soumis pour publication et au moins une communication dans une conférence internationale sera assurée. Par ailleurs, une fois que l’algorithme de sélection des cas (décrit ci-dessous) sera validé et publié, il sera mis à disposition de la communauté scientifique en open source pour une éventuelle réutilisation. Il sera également référencé sur la [documentation ouverte](https://documentation-snds.health-data-hub.fr/) du SNDS.

1. **Méthodologie [3 - 5 pages maximum]**
   1. **Design de l’étude**

Il s’agit d’une étude de cohorte rétrospective avec une analyse longitudinale sur une période d’étude comprise entre le 01/01/2010 et le 31/12/2019. Une cohorte de patients suivis sur une aussi longue période nous permettra d’évaluer la prévalence annuelle de la maladie et son évolution entre 2010 et 2015 . Cette analyse nous permettra également de réaliser le suivi de leur prise en charge (2010-2019). A notre connaissance, aucune donnée récente n'est disponible sur la prévalence de l’EGT1 en France.

Deux cohortes seront mises en place. Cohorte 1 : mise en place à partir des données de la base principale du SNDS. Cohorte 2 : issues de l’appariement indirect entre l’étude clinique Beta et les données de la base prinicpale du SNDS. Cet appariement est nécessaire car le diagnostic de l’EGT1 dans l’étude clinique Beta est confirmé par un médecin (données non disponibles dans la base principale du SNDS). Par ailleurs, la cohorte 2 nous permettra de valider l’algorithme de sélection de cas qu’on aura élaboré avec la cohorte 1.

Le critère principal de l’étude sera la prévalence de l’EGT1 et les critères secondaires seront les ressources consommées et les coûts des soins ainsi que le délai entre la date des premières manifestations patentes de troubles neurologiques et la date du diagnostic de l’EGT1.

* 1. **Description et justification de la population d’étude**

***Cohorte 1 :***

La population d’étude sera constituée de l’ensemble des individus atteints de l’EGT1 en France entre 2010 et 2015. Il seront sélectionnés par la mise en place d’un algorithme de sélection des cas appliqué sur les données de la base principale du SNDS. Cet algorithme sera validé en comparant les résultats avec ceux de la cohorte 2 (définis ci-dessous).

*Pourquoi un algorithme de sélection des cas ?*

Le code diagnostic (CIM-10) de l’EGT1 est le G93.4 (Encéphalopathie, sans précision). Ce code n’est pas spécifique de cette affection puisqu’il englobe d’autres formes d’encéphalopathies (référence bibliographique). Afin d’identifier les patients atteints spécifiquement de l’EGT1 dans la base principale du SNDS, un algorithme de sélection est donc nécessaire.

Les critères d’inclusion des patients seront :

* **[**Les patients avec une affection de longue durée (ALD) active entre le 01/01/2010 et le 31/12/2015 avec un diagnostic d’encéphalopathie (code CIM-10 G93.4) ;

**OU**

Les patients hospitalisés au moins une fois en MCO, HAD, SSR ou RIM-P entre le 01/01/2010 et le 31/12/2015 avec un DP, DR ou DA d’encéphalopathie (code CIM-10 G93.4)];

**ET**

* Les patients avec au moins un examen cytobacteriologique d'un liquide céphalorachidien (code NABM 5231) **ET** un dosage de la glycémie (code NABM 0552) entre le 01/01/2010 et le 31/12/2015.

Le ciblage des patients portera sur la période allant de 2010 à 2015 pour permettre l’estimation de la prévalence annuelle. L’extraction des données portera sur les années 2010 à 2019 afin de suivre leur parcours de soins. Concernant les composantes de la base principale du SNDS, le DCIR et le PMSI ainsi que les tables des référentiels (Bénéficiaire (table IR\_BEN\_R), le référentiel médicalisé des bénéficiaires (table IR\_IMB\_R) et Pharmacie (table IR\_PHA\_R)) seront nécessaires. Concernant le PMSI, du fait que ce soit une maladie rare, tous les champs seront nécessaires pour l’extraction et le ciblage des données (MCO, HAD, SSR et RIM-P). Cependant, les tables des causes de décès du CépiDc ne seront pas utilisées. Les dates de décès présentes à la fois dans le DCIR, le PMSI ainsi que dans la table du référentiel des bénéficiaires permettront de réaliser les analyses.

***Cohorte 2 :***

Dans cette deuxième cohorte, la population d’étude sera composée des 115 patients résidents en France atteints de l’EGT1 inclus dans l’étude clinique Beta. Le recueil des données de l’étude Beta avait été effectué entre le 01/01/2014 et le 31/12/2014.

Appariement et extraction : L’appariement des données sera donc effectué entre le 01/01/2014 et le 31/12/201ça mar4. Une fois les données appariées, une extraction de 2010 à 2019 sera nécessaire. Cette longue période d’extraction nous permettra de réaliser un suivi longitudinal des patients et de valider l’algorithme de sélection des cas mis en place dans la cohorte 1. Les mêmes champs décrits pour la cohorte 1 seront utilisés.

* 1. **Taille de la population**

**Cohorte 1 :**

Concernant la taille de la population, à titre indicatif, 9784 hospitalisations pour Encéphalopathie (code CIM10 G93.4 en DP uniquement) ont été observées en 2019. Ce chiffre a été obtenu à partir des données open source de l’ATIH (source : <https://www.xx>). De plus, dans une précédente étude, la prévalence de l’EGT1 a été estimée à 972 cas en 2010 en France métropolitaine (référence bibliographique). A partir de ces éléments, nous estimons que la taille de la population finale pour cette cohorte sera composée d’environ 5 000 individus sur toute la période d’extraction de 2010 à 2019.

**Cohorte 2 :**

La cohorte 2 sera composée de la population d’étude de l’étude clinique Beta après un appariement indirect avec les données de la base principale du SNDS. Selon le pourcentage d’appariement, la population d’étude sera composée de 115 patients à laquelle on va soustraire le nombre d’individus pour lesquels l’appariement n’aura pas permis leur identification. Ces patients seront suivis de 2010 à 2019.

* 1. **Source de données**

Deux sources de données seront exploitées dans cette étude:

* **Source 1 : Les données de la base principale du SNDS :**

C’est une base médico-administrative qui contient l’ensemble des informations se rapportant aux dépenses remboursées par les différents régimes d’assurance maladie (Tuppin et al. 2017). Elle contient également différentes données de nature démographique (âge, sexe, statut vital), ainsi que pour les personnes exonérées du ticket modérateur, les diagnostics des affections de longue durée (ALD), codés selon la classification internationale des maladies (CIM-10). Ces données issues du Système national d'information inter-régimes de l'Assurance maladie (SNIIRAM) sont chainées aux données du programme de médicalisation des systèmes d'information (PMSI) soit les séjours des patients hospitalisés dans les établissements publics et privés d’hospitalisation, diagnostics de sortie et actes médicaux réalisés durant le séjour. Les données sont également appariées à la base du Centre d'épidémiologie sur les causes médicales de Décès (CépiDc) de l’Inserm, pour les causes médicales de décès. Les données des bases relatives à l’épidémie de covid 19 (bases Vaccin-covid et SI-DEP) alimentent également la base principale du SNDS.

Ces données de la base principale du SNDS sont indispensables pour répondre aux objectifs de notre étude : pour l’étude de la prévalence de l’EGT1 dans toutes les régions de France de 2010 à 2015, du parcours de soins en prenant en compte les soins ambulatoires et hospitaliers jusqu’au décès, ainsi que l’utilisation des ressources en santé et leur coûts associés.

* **Source 2 : Les données de l’étude clinique Beta :**

L’étude clinique Beta a été autorisée par la CNIL depuis le 31 décembre 2013 (autorisation n°19123456, délibération 2017-01234). Elle s’est achevée le 31/12/2014 et les résultats ont été publiés récemment (référence bibliographique). Brièvement, la population d’étude est composée de 115 patients recrutés par le médecin investigateur de l’étude dans les centres de références sur la base du volontariat entre le 01/01/2014 et le 31/12/2014. Des eCRF ont été générés et la base de données a été consolidée le 30/06/2015. L’eCRF est détaillé en annexe.

Telle que décrite précédemment, la cohorte 2 sera construite par la mise en place d’un appariement indirect des données de l’étude Beta avec les données de la base principale du SNDS. La circulation des données pour réaliser l’appariement est décrite dans le chapitre “Circuit des données en cas d’appariement”.

Les variables que nous avons identifiées pour l’appariement indirect sont : le mois et l’année de naissance ; le sexe ; le code FINESS de l’hôpital ; la date d’hospitalisation (date d’entrée et de sortie) ; le diagnostic de la maladie (l’encéphalopathie, code CIM-10 G93.4) ; les examens biologiques effectués (codes NABM 5231 et 0552) la date de l’examen biologique (code NABM 5231 et 0552).

Le mois et l’année de naissance (traités ensemble) sont considérés comme données sensibles. Toutefois, ils nous seront indispensables pour augmenter la probabilité de réussite de l’appariement. En effet, nous ne disposons pas d’autres variables plus discriminantes.

L’appariement nous permettra d’étudier le parcours de soins, les ressources en soins de santé ainsi que les coûts associés. De plus, nous utiliserons les données appariées pour la validation de l’algorithme développé avec les données de la cohorte 1.

* 1. **Variables**

*Les informations rapportées dans le tableau qui suit sont fictives et données à titre d’exemple. Il ne s’agit en aucun cas de données exhaustives.*

Les variables qui seront utilisées dans le cadre de ce projet sont détaillées dans le tableau ci-dessous :

|  |  |  |
| --- | --- | --- |
|  | **Définition** | **Type** |
| **Critères** |  |  |
| Critère principal |  |  |
| La prévalence de l’EGT1 | Proportion de patients diagnostiqués avec l’EGT1, par an de 2010 à 2019 | N / % |
| Critères secondaires |  |  |
| Hospitalisation | au moins une date d’hospitalisation renseignée | Binaire (oui/non) |
| Nombre d’hospitalisation | la somme des hospitalisations | Continue |
| Durée du séjour à l'hôpital | Date de sortie - date d’hospitalisation | Continue / en jours |
| Les coûts | le montant des coûts associés à l'utilisation des ressources de santé par personne-année. | Continue / en Euros |
| Le délai entre la première manifestation de troubles neurologiques et la date du diagnostic EGT1 sera calculé. | Date du premier diagnostic de l’EGT1 moins la date de première manifestation de troubles neurologiques | Continue / en jours |
| **Description de la population** |  |  |
| Âge (continue) | / | Continue |
| Sexe | / | Homme/femme |

* 1. **Préparation des données**

Avant de procéder aux analyses de données, un travail préalable sera effectué sur la préparation et la mise en qualité des données. L'évaluation de la disponibilité et de l'exhaustivité des variables, y compris les données manquantes sera effectuée.

**Cohorte 1 :**

S’agissant de données issues de la base principale du SNDS, et bien qu’on ne s'attende pas à avoir beaucoup de valeurs manquantes et/ou aberrantes, elles seront également analysées et documentées.

**Cohorte 2 :**

Les données permettant de faire l’appariement entre l’étude clinique Beta et le SNDS seront préparées et mises en qualité avant leur transmission pour appariement.

Les données recueillies auprès des patients et de leur médecin investigateur dans le cadre de l’étude clinique Beta (eCRF en annexe) ont été pseudonymisées et un numéro de patient a été généré aléatoirement. Du fait que nous ne disposons pas du NIR des patients ni des variables nécessaires pour sa reconstitution, un appariement indirect sera réalisé. Une table d’appariement intégrant le numéro d’appariement et les variables communes entre les deux tables a été constituée.

* 1. **Méthodes, traitements et analyses des données**

Toutes les analyses statistiques seront effectuées au moyen du logiciel SAS, version 9.4 ou ultérieure.

Des analyses descriptives, univariées et multivariées seront effectuées. Tous les tests statistiques seront bilatéraux et considérés comme significatifs à un niveau de 5 %.

**Analyses descriptives**

Les caractéristiques démographiques et cliniques des patients seront décrites en calculant la fréquence et le pourcentage pour les variables catégorielles, et moyennes, médiane, écarts interquartiles et écart-type pour les variables continues. Les valeurs aberrantes seront incluses dans la description.

Concernant la prévalence de l’EGT1, elle sera estimée annuellement à partir des données du SNDS de 2010 à 2015.

L'utilisation d'une longue période permettra de voir l'évolution du nombre de cas prévalents. Des taux de prévalence standardisés sur l'âge et le sexe pour 100 000 habitants, selon la population européenne (données Eurostat) seront calculés.

Afin de valider l’algorithme décrit en paragraphe 3.2, les résultats obtenus à partir des données de la base principale du SNDS (cohorte 1) seront comparés à ceux obtenus après appariement de l’étude clinique Beta (cohorte 2). La sensibilité, la spécificité, la valeur prédictive positive et négative, VPP et VPN, seront estimées. Des variations dans les indicateurs de l’algorithme pourront être effectuées afin d’améliorer sa performance.

Les ressources de soins consommées ainsi que leurs coûts selon les perspectives sociétales et assurance maladie portant sur la période 2010-2019 feront l’objet d’une analyse descriptive. Les coûts seront rapportés, de façon annuelle, par grands postes de dépenses (hospitaliers, consultations, actes et visites de médecins, actes des paramédicaux, médicaments achetés en officine). Le délai entre la date des premières manifestations patentes de troubles neurologiques identifiables dans la base et la date du diagnostic EGT1 sera calculé.

**Analyses comparatives**

Des analyses comparatives seront effectuées à partir des données de la cohorte 2 (données appariées). Les patients atteints de l’EGT1 ayant subi au moins un acte spécifique de l’affection seront comparés à ceux n’ayant jamais subi d’actes spécifiques de l’affection. Pour les variables d'intérêts qualitatives, le test du Chi² de Pearson sera utilisé à moins que l’effectif théorique estimé dans une cellule soit inférieur à cinq, auquel cas, la correction de continuité de Yates ou le test exact de Fisher seront appliqués. Pour les variables quantitatives, le test t ou une analyse de la variance seront effectués. Sinon, les analyses post hoc seront effectuées à l'aide du test de Student-Neuman-Keuls. Si les données ne sont pas distribuées normalement, des tests non paramétriques seront utilisés.

Les analyses comparatives incluront:

* les différences dans l’utilisation des ressources de soins chez les patients atteints de EGT1 ayant subi versus n’ayant pas subi d’actes spécifiques de l’affection ;
* les différences des coûts chez les patients atteints de EGT1 ayant subi versus n’ayant pas subi d’actes spécifiques de l’affection.
  1. **Limites de l’étude**

S'agissant d’une maladie rare, les analyses par sous-groupes dans la cohorte 2 peuvent être limitées du fait du nombre insuffisant de sujets nécessaires. C’est pour cela que la validation de l’algorithme nous permettra d’identifier les cas à partir de toute la population française. Comme dans toutes les bases de données médico-administratives, les erreurs de codage des diagnostics dans le SNDS peuvent entraîner des biais de classifications. En milieu hospitalier, les médicaments sont inclus dans les tarifs des séjours hospitaliers et ne sont donc pas repérables dans la base principale du SNDS (PMSI). Les seuls repérables sont les médicaments innovants et coûteux faisant partie de la liste en sus.

L'estimation de la prévalence pourrait varier en fonction de la région de résidence, de la démographie médicale et de l'accès aux soins (accès difficile aux spécialistes dans les territoires où le nombre de neurologues par habitant est faible). L’utilisation des données cliniques de la cohorte 2 après appariement nous permettra de répondre à un grand nombre de limites citées ci-dessus.

* 1. **Calendrier prévisionnel et faisabilité du projet**

**Calendrier prévisionnel**

* + Avis favorable initial du CPP pour l’étude clinique Beta: 30/06/2020.
  + Rédaction du protocole scientifique et sa validation par le conseil scientifique de l’étude : 31/01/2021.
  + Avis CESREES : ~ 28/02/2021. et Avis de la CNIL : ~ 30/04/2021.
  + Signature de la convention : ~ 30/06/2021.
  + Accès aux données : ~ 31/07/2021.
  + Analyses des données et validation des résultats par le conseil scientifique de l’étude : ~ entre le 31/07/2021 et le 31/10/2021.
  + Rédaction et validation du rapport scientifique : ~ 31/12/2021.
  + Publication des résultats : ~2022/2023 (des analyses supplémentaires pourraient être demandées par les reviewers des publications scientifiques).

**Faisabilité du projet**

Le président de l’organisme responsable de traitement est docteur en Médecine Vétérinaire de l'Ecole Nationale Vétérinaire complété par un doctorat en biologie cellulaire et microbiologie de l'Institut Pasteur et de l'Université Paris VII (Paris, France).

Le responsable de mise en œuvre de cette étude, le bureau d’étude HealthX a déjà été responsable de la mise en œuvre de 12 études utilisant les données du SNDS. L'équipe projet a une bonne expérience en traitement des données qui seront analysées dans ce projet. Un datamanager et un biostatisticien expérimentés assureront les analyses des données. Ils ont tous les deux effectués les formations “Données d’extraction DCIR via un espace projet” et “EGB” et ont une expérience reconnue sur les analyses des données médico-administratives (y compris les données du SNDS) ainsi qu’en phamaco-épidémiologie.

Un comité scientifique indépendant a été mis en place dans le cadre de ce projet, Pr XX, MD, biostatisticien et pharmaco-épidémiologiste ; Pr XX, MD, neurologue ; Pr XX, économiste de la santé et Mme XX, représentante d’une association d’usager du système de santé.

Le délégué à la protection des données, est également Directrice des Affaires Réglementaires et Responsable de l’Assurance Qualité, au sein de l’organisation PharmaX . Il a une double expérience professionnelle en biologie et gestion de la qualité. Elle est titulaire d'un diplôme en biologie cellulaire, spécialisée en recherche biomédicale et immunologie et d'un diplôme en environnement de santé et sécurité de qualité (QHSE).

Le projet est financé à 100 % par le laboratoire pharmaceutique PharmaX.

Les délais de réalisation de ce projet sont compatibles avec l’état d’avancement de l’étude.

# Protection de la vie privée, sécurité et confidentialité des données [10 PAGES MAXIMUM]

(Section destinée à la CNIL)

* 1. **Respect des droits des personnes concernées**

**S’agissant de l’information :**

Les patients atteints d’EGT1 dont les données seront utilisées pour cette étude ont signé deux types de consentements. Le premier avant le début de l’étude clinique Beta pour l’utilisation de leur données personnelles et un deuxième consentement qu’ils ont signé *a posteriori*, après avoir reçu une note d’information concernant l’utilisation de leurs données de la sécurité sociale (SNDS). Ces documents sont joints en annexes.

S’agissant d’une population d’étude composée principalement par des mineurs, nous avions demandé la signature d’au moins un représentant légal. Les enfants avaient également eus une lettre d'information dont le contenu a été expliqué par l'investigateur.

**Pour l’exercice des droit plus généralement :**

Concernant les données provenant de la base de données du Système National de l’Information Inter-Régimes de l’Assurance Maladie (SNIIRAM) incluse dans la base principale du SNDS, les personnes concernées souhaitant faire valoir leurs droits devront en faire la demande auprès de leur CPAM de rattachement, conformément à l’article 111 du décret d’application de la loi Informatique et Libertés. Cette procédure est détaillée sur le site <https://www.snds.gouv.fr/SNDS/Protection-de-la-> [donnee](https://www.snds.gouv.fr/SNDS/Protection-de-la-donnee).

Pour les données provenant de l’étude clinique Beta, les patients pourront faire valoir leurs droits auprès de leur médecin investigateur. Le médecin pourra ensuite adresser les questions au Délégué à la Protection des Données du Promoteur en utilisant le code d'identification unique du patient pour protéger sa confidentialité.

* 1. **Confidentialité et sécurité des données**
     1. **Gestion du risque de réidentification**

Les données de la base principale du SNDS sont pseudonymisées et ne contiennent aucune donnée directement identifiante. L’identifiant patient utilisé est dérivé du NIR de manière irréversible par applications successives des différentes fonctions de pseudonymisation FOIN.

Parmi les variables indirectement identifiantes, les suivantes sont généralisées :

* + La date de naissance est tronquée pour ne conserver que l’année et le mois
  + L’adresse de résidence est généralisée au code commune

Les données de la base principale du SNDS ne contiennent aucune donnée liée à l’affiliation ethnique, politique, religieuse, philosophique ou syndicale, ni donnée génétique, ni donnée biométrique, ni donnée sur la vie ou l’orientation sexuelle.

**Durée de conservation des données**

La base appariée avec les données extraites de la base principale du SNDS sera conservée pendant une durée de 3 ans sur la plateforme du Health Data Hub. Seules seront extraites de cette plateforme, des données agrégées sous forme de tableau ou de graphique (résultats de l’étude).

La base de données qui servira pour l’appariement sera détruite dès que l’appariement sera validé et les données intégrées dans la plateforme.

* + 1. **Support des données**

L’analyse des données aura lieu sur un espace projet dédié, hébergé sur la plateforme technologique du Health Data Hub, homologuée au référentiel de sécurité SNDS. Des mesures de sécurité s’appliquant à cet espace projet ont été mises en place à la fois par le Health Data Hub, conformément à la démarche de sécurité validée par l’ANSSI et la CNIL et dont les détails ont été transmis à ces institutions, et par l’équipe projet, conformément au dossier d’homologation de l’espace projet qui est joint à ce protocole dans la demande d’autorisation.

Toutes les analyses des deux cohortes 1 et 2 se feront sur l’espace projet sécurisé de la plateforme du Health Data Hub. Accessibles aux personnes habilitées seulement. Aucune donnée individuelle ne sera enregistrée en dehors de la plateforme. Seuls les résultats anonymes seront exportés sous forme de tableaux et de graphiques. L’accès à la plateforme se fera par un data manager et un statisticien expérimentés, tous deux habilités à accéder aux données de la base principale du SNDS ainsi qu’à la plateforme du Health Data Hub.

Le recueil de données de l’étude clinique Beta qui va être apparié au SNDS a été réalisé dans le respect des bonnes pratiques cliniques et a été soumis et accepté par le CPP (n°xxxx).

* + 1. **Circuit des données en cas d’appariement**

Les données recueillies auprès des patients et de leur médecin investigateur dans le cadre de l’étude clinique Beta ont été anonymisées et un numéro de patient a été généré aléatoirement. Une table d’appariement intégrant le numéro d’appariement et les variables communes entre les deux tables a été constituée.

Tel que préconisé par la Cnam, cette table d’appariement sera chiffrée avec GnuPG et déposée sur leur plateforme sécurisée Petra.

Nous transmettrons ensuite les données anonymisées de notre étude clinique Beta au Health Data Hub. L’extraction de données issues de la base principale du SNDS de notre population d’étude sera également transmise au Health Data Hub par la Cnam après application du ciblage à partir de l’algorithme décrit dans le protocole scientifique (chapitre 3.1.1). Ces transmissions de données se feront selon les préconisations du Health Data Hub et de la Cnam via des réseaux dédiés, permettant un transfert privé (hors internet) et sécurisé des données.

# Annexes

* Bibliographie
* Les eCRF de l’étude clinique Beta
* Notice d’information soumise aux patients ou aux personnes concernées
* Formulaire de consentement
* Liste des variables utilisées

**Annexe 2 – Expression de besoins**

**données de la base principale du SNDS**

A compléter si votre projet nécessite une extraction de données de la base principale du SNDS,

Ce formulaire doit obligatoirement être rédigé en français

**Le périmètre de l’extraction sera établi sur la seule base de ce formulaire et de l’autorisation CNIL**

**Création : JJ/MM/AAAA**

**Version : JJ/MM/AAAA**

**Statut :** ☐ **Soumis CESREES /** ☐ **Soumis CNIL**

|  |  |  |
| --- | --- | --- |
| Spécifications générales de la demande :   * **Titre du projet** : [Nom] * **Version du protocole**: [Nom] et [Date] * **Description des populations faisant l’objet de la mise à disposition par la Cnam**: (population(s) à extraire par la Cnam par ciblage ou appariement, population témoin, intersection entre les populations…) * **Type de traitement :**   ☐ Nouveau traitement  ☐ Modification du traitement initial (dans le cadre d’une modification substantielle) | | |
| **Acronyme du projet : [Acronyme]** | N° de dossier Démarche simplifiée  :  Date de l’avis du CESREES :  Numéro de demande CNIL : [N°]  N° et date de l’autorisation CNIL : [N°], [date]  ou - Autorisation tacite : [date] | |
| L’identité du ou des responsables du traitement (RT) | **Responsable(s) du traitement :**  Organisme ou Société :  Adresse postale :  Représenté par :  Nom, Prénom :  Fonction :  Adresse mail :  N° Tél : | |
| L’identité de l’éventuel responsable de la mise en œuvre du traitement de données (RMOT) – *laboratoire de recherche ou bureau d’études* | **Responsable de la mise en œuvre du traitement :**  Organisme ou Société :  Adresse postale :  Représenté par :  Nom, Prénom :  Fonction :  Adresse mail :  N° Tél : | |
| L’identité du contact opérationnel | **Contact opérationnel :**  Nom, Prénom :  Organisme ou Société :  Adresse mail :  N° Tél : | |
| Autorisation de traitement des données | Composantes de la base principale du SNDS autorisées et années d’extraction autorisées :  ☐ DCIR (Assurance Maladie), Années :  ☐ PMSI (ATIH), Années :  ☐ Causes médicales de décès (CépiDc), Années :  Durée d’accès aux données autorisée : | |
| Fréquence de traitement | ☐ Unique  ☐ Plurielle : préciser la fréquence : [Fréquence] | |
| ☐ Ciblage  Et/ou  ☐ Appariement | En cas de ciblage :   * Critères de ciblage par population à mettre en œuvre ( (préciser : régimes AM, code(s) CIM10, tranches d’âges, code(s) ATC/CIP, périmètre géographique,…) : * Période de ciblage : * Volume estimé (Nombre d’individus/de cas ciblés, par population le cas échéant) : | |
| En cas d’appariement :  ☐ Projet monocentrique  ☐ Projet multicentrique - Préciser le nombre de centres participants : xxx  ☐ Appariement indirect  Effectif de la population à apparier :  ☐ Appariement direct :  ☐ Via l’utilisation du NIR (préciser le référent technique autorisé à transmettre les NIR ci-dessous) :  ☐ Avec reconstitution du NIR (préciser le référent technique autorisé à transmettre les données d’état civil ci-dessous) :  Nom, Prénom :  Organisme ou Société :  Adresse mail :  N° Tél :  Effectif de la population à apparier : | |
| Identifiants potentiels souhaités | ☐ Année et mois de naissance  ☐ Date de soin (jour + mois + année)  ☐ Date du décès (jour + mois + année)  ☐ Commune de résidence  ☐ Commune de décès | |
| Champ d’extraction et périodes à extraire | ☐ Données de consommations individuelles (DCIR), Période :  ☐ Référentiel médicalisé (IR\_IMB\_R) (la période sera alignée sur celle de DCIR)  Données d’hospitalisation (PMSI, champs à préciser), Période :  ☐ PMSI-MCO  ☐ PMSI-HAD  ☐ PMSI-SSR  ☐ PMSI-RIM-P  ☐ Causes de décès, Période :  ☐ Autres :  ☐ Référentiel Pharmacie (IR\_PHA\_R) | |
| **Modalités de restitution** | | |
| ☐ Espace projet du HDH  ☐ Système fils du SNDS  **Gestionnaire du système fils**  Organisme ou Société :  Adresse postale :  Représenté par :  Nom, Prénom :  Fonction :  Adresse mail :  N° Tél :  **Personne référente pour la réception des données**  Nom, Prénom :  Adresse mail :  N° Tél :  **Modalités de restitution des données :**  ☐ Disque dur (capacité en fonction de la volumétrie)  Chiffrement des données à partir d’une clé GnuPg  ☐ Plate-forme de téléchargement sécurisé (**compatible Cnam**)  **Format des Données restituées :**  ☐ SAS  ☐ CSV  \**homologué au référentiel de sécurité* | | ☐ Espace projet sous le portail SNDS  Dans un libname SAS via le profil SNDS n°117 :  DMX**xxxxx \***  Espace projet accessible aux seules personnes habilitées  \**DMX suivi de 5 caractères maximum* |

1. Le SNDS rassemble l’ensemble des données de santé financées par la solidarité nationale [telles que visées par le Code de Santé Publique](https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006072665/LEGISCTA000031923877/#LEGISCTA000031923880). Le SNDS dit “central” comprend notamment la base principale du SNDS qui se compose des données suivantes :

   * Datamart de consommation inter régimes (DCIR) rassemblant les données ambulatoires ;
   * La base PMSI en provenance des établissements de santé :
     + des données sur les champs médecine, chirurgie, obstétrique et odontologie (MCO)
     + des données de soins de suite et de réadaptation (SSR)
     + du recueil d’information médicalisée en psychiatrie (RIM-P) et hospitalisation à domicile (HAD)
   * La base de causes médicales de décès (base du CépiDc de l’Inserm)
   * Les données relatives au handicap (données de la CNSA)
   * Les données des bases relatives à l’épidémie de covid 19 (bases Vaccin-covid et SI-DEP)

   [↑](#footnote-ref-1)