Bac à sable « santé numérique »

Les recommandations de la CNIL aux lauréats

À noter

Ce document constitue une **synthèse des principales recommandations** formulées au porteur de projet lors de son accompagnement « bac à sable ». Ces recommandations, applicables à la date de rédaction du document, s'appuient sur les informations communiquées par le porteur de projet et ses échanges avec la CNIL.

Elles sont publiées pour en faire **bénéficier les acteurs du secteur de la santé numérique** et **aider les innovateurs sur des projets similaires** à développer leur solution

2022

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

Table des matières

Le projet du CHU de Lille et de l'équipe Magnet de l'Inria : mettre en œuvre un procédé d'apprentissage fédéré entre plusieurs entrepôts de données de santé

Date de rédaction : février 2022

De juin à décembre 2021, la CNIL a accompagné le CHU de Lille, en coopération avec d'autres centres hospitaliers de la région. Cet accompagnement s'est concentré sur la mise en place d'un **protocole d'apprentissage fédéré** portant sur un **algorithme** visant à faciliter la prise en charge des patients et utilisant des données hébergées dans plusieurs entrepôts de données de santé (EDS).

Accompagné par l'équipe de recherche Magnet d'Inria, le CHU a mis en œuvre les recommandations de la CNIL sur deux projets de recherche :

- un projet d'aide au codage des diagnostics médicaux par analyse textuelle (projet *Audimat*)[1] ;
- un projet de prédiction automatisée de la ré-hospitalisation par apprentissage décentralisé (projet *Parade*)[2].

S'agissant du **contexte juridique**, les données des entrepôts de données de santé (EDS) concernés, autorisés par la CNIL, ont pu être réutilisées dans le cadre d'une recherche n'impliquant pas la personne humaine (RNIPH), sous réserve du respect de certaines exigences réglementaires.

Le protocole, en ce qu'il utilise des techniques innovantes d'intelligence artificielle, a été considéré comme un projet de recherche dans le domaine de la santé, et sa réalisation est donc soumise au respect des formalités applicables[3].

En ce qui concerne l'**architecture technique**, les données utiles hébergées au sein des EDS étaient prétraitées et utilisées pour l'entraînement au sein de l'infrastructure de chaque EDS uniquement.

Seuls les agrégats et paramètres utilisés au cours de l'apprentissage étaient extraits des EDS lors des itérations de l'apprentissage le nécessitant, pour être agrégés au sein d'un centre orchestrateur également situé au sein de l'infrastructure informatique d'un EDS.

La technique de l'apprentissage fédéré, qui permet de **développer un modèle** d'intelligence artificielle **sans centraliser les données,** a soulevé plusieurs questions auxquelles la CNIL a répondu lors de son accompagnement. Cet accompagnement a également permis à la CNIL de renforcer son expérience en la matière, en vérifiant notamment les apports de l'apprentissage fédéré comme **technique protectrice de la vie privée**.

Les trois questions suivantes ont été traitées lors de l'accompagnement « bac à sable ».

1. Évaluer la nature des agrégats

Établir la nature des agrégats – données personnelles ou anonymes – issus des itérations de l'apprentissage est une étape cruciale en ce qu'elle permet de déterminer le régime juridique applicable au traitement de données lors de leur export d'un EDS.

Afin d'apprécier cette nature, le responsable de traitement (RT) a été encouragé à **vérifier le caractère anonyme des données le plus en amont possible dans le traitement envisagé**, un résultat issu de l'agrégation de données anonymes pouvant être considéré comme anonyme à son tour.

Lorsque le caractère anonyme des données d'entraînement ne peut être vérifié, une **analyse des risques de réidentification portant sur les agrégats** finaux ou utilisés au cours de

[1] Décision DR-2022-171 du 26 juillet 2022
[2] Décision DR-2022-169 du 26 juillet 2022
[3] Pour un projet de recherche dans le domaine de la santé, le responsable de traitement doit obtenir une autorisation préalable de la CNIL, si son étude n'est pas conforme à une méthodologie de référence.

l'apprentissage doit être réalisée. A cet égard, le choix d'un algorithme « explicable » (i.e. dont le fonctionnement général est connu et dont les décisions individuelles peuvent être justifiées) et utilisant un faible nombre de paramètres peut faciliter cette analyse de risque.

De même, tout pré-traitement des données d'entraînement permettant d'en limiter le potentiel d'identification, tel que l'*embedding*[4] ou l'ajout de bruit, peut être mobilisé. Toutefois, si cette analyse de risque donnait des résultats insuffisants, une position de précaution devrait être adoptée, et **les agrégats issus de l'apprentissage ou utilisés lors de l'apprentissage devront être considérés comme des données personnelles**.

2. Déterminer le cadre juridique applicable à l'export de données non-anonymes d'un entrepôt de données de santé

Lorsque les agrégats ne peuvent être considérés comme des données anonymes dans un apprentissage fédéré, il en résulte qu'un export de données personnelles a lieu lors de l'entraînement de l'algorithme.

Dans le cas d'espèce, la CNIL a considéré au vu de leur rédaction que **les autorisation EDS délivrées permettaient un tel export**, sous réserve de l'accomplissement de formalités préalables (demande d'autorisation recherche spécifique), et du respect de certaines modalités d'information des personnes et d'exercice des droits, notamment.

3. Mettre en place les mesures techniques appropriées afin de sécuriser le traitement

Dès lors que des données non anonymes sont exportées des EDS, une analyse *ad hoc* des risques doit être effectuée afin de déterminer les mesures de sécurité nécessaires. Il convient toutefois d'assurer autant que possible la continuité des mesures de sécurité en dehors de l'entrepôt. Pour cela, le RT pourra mettre en place les mesures de sécurité habituelles telles que le chiffrement TLS des communications (utilisé par le protocole HTTPS), mais **également des mesures de sécurités propres à l'apprentissage fédéré**, telles que le chiffrement homomorphe (une méthode permettant de réaliser des opérations sans perte de confidentialité des données chiffrées).

[4]Plongement (*embedding*) : représentation logique de données permettant d'en extraire uniquement les informations pertinentes.

Le projet de Resilience : une solution d'aide au diagnostic en oncologie

Date de rédaction : avril 2022

En 2021, la CNIL a accompagné Resilience souhaitant proposer une application mobile d'éducation thérapeutique et un logiciel d'aide à la prise de décision médicale. Ce dernier est constitué d'un algorithme d'intelligence artificielle développé grâce à un entrepôt de données de santé, comportant notamment les données issues de l'application mobile.

La start-up Resilience a débuté avec une **application mobile d'éducation thérapeutique** pour les patients atteints de cancer. Par la suite, Resilience a développé **un logiciel d'aide à la décision médicale** fonctionnant grâce à un algorithme d'intelligence artificielle.

Pour réunir ces deux outils, Resilience a souhaité créer un **entrepôt de données de santé** notamment alimenté par des données collectées via son application mobile et des données conservées au sein d'entrepôts hospitaliers d'établissements partenaires.

Au cours de son accompagnement, la jeune pousse a racheté une **société de télésurveillance** : les données issues de ce type de suivi ont donc été ajoutées aux deux premiers jeux de données.

La CNIL a travaillé avec le délégué à la protection des données de Resilience et ses équipes afin de mettre en place un **entrepôt de données de santé** poursuivant une finalité d'intérêt public.

Ces échanges ont permis d'organiser **l'interconnexion de différentes sources** ainsi que **la réutilisation des données rassemblées**.

Les trois questions suivantes ont été abordées lors de l'accompagnement « Bac à sable ».

1. Comment rassembler les données de différentes sources pour constituer un entrepôt ?

L'objectif du projet de *Resilience* est notamment de collecter et des données personnelles issues de différentes sources (applications mobile ou web, outil de télésurveillance, entrepôts de données de santé préexistants, dossiers patients informatisés, *etc.*)[5].

Centraliser ces différentes informations et données personnelles, notamment des données de santé, correspond à un **entrepôt de données de santé**. Il s'agit d'un traitement soumis au RGPD et à la loi « informatique et libertés » modifiée.

Pour déployer un tel entrepôt de données de santé, le responsable est tenu à plusieurs obligations, et notamment :

- **Définir une finalité**, c'est-à-dire l'objectif pour lequel les données sont collectées et utilisées. Cette finalité devra notamment figurer dans l'information délivrée aux personnes pour leur permettre de comprendre à quoi vont servir leurs données.

> Le projet de *Resilience* a pour la finalité la constitution d'un entrepôt pour mener des études n'impliquant pas la personne humaine dans le domaine de la cancérologie.

- **Établir une base légale :** sur ce point, voir la question n° 2 du document (« Comment articuler le concept d'intérêt public avec un traitement s'inscrivant dans un cadre commercial ? »

[5] Ces données sont complétées avec d'autres informations librement accessibles issues de la littérature scientifique et médicales.

- **Vérifier la source des données :** le responsable doit s'assurer que les données personnelles réutilisées ont été initialement collectées et conservées dans le cadre d'un **traitement régulièrement constitué**.

Il s'agit notamment de vérifier la durée de conservation des données (est-elle non dépassée par le responsable de traitement initial ?), les modalités d'information initiales des personnes (les personnes ont-elles été bien informées ?), les destinataires (qui sont les destinataires identifiés à l'origine ?), la profondeur historique des données, *etc.*
Si le traitement initial était soumis à une formalité préalable, il est aussi nécessaire de vérifier que cette formalité bien été accomplie auprès de la CNIL.

- **Minimiser les données :** c'est-à-dire collecter uniquement les données strictement nécessaires par rapport à l'objectif de cette collecte.

Pour le projet de *Resilience*, cette minimisation implique notamment une **pseudonymisation** des données à la source, c'est-à-dire par le responsable de traitement initial. Il peut toujours être démontré sur la base d'éléments de contexte précis que cette opération n'est pas envisageable ou possible.

- **Définir une durée de conservation** pour chaque catégorie de données et en fonction des finalités poursuivies.

Dans le cas de *Resilience,* une durée de quinze ans a été retenue pour les données de santé. Il est également prévu de mettre en place des alertes automatisées en cas de dépassement de cette durée.

- **Informer les personnes concernées**

Pour informer correctement les personnes concernées (principe de transparence[6]), **l'information doit être délivrée à différents stades et par différents responsables de traitement** :

- par les responsables de traitement **initiaux** concernant la réutilisation des données conservées dans leur base ;
- par **chaque responsable de traitement** procédant à **la réutilisation de données** précédemment collectées par un autre responsable de traitement. Dans le cas présent, *Reslience* est tenue d'informer les personnes concernées de la mise en œuvre de son entrepôt ainsi que de toute étude qu'elle met en œuvre.

Pour un traitement de données dans le domaine de la santé[7] (tel que l'entrepôt de données de santé de *Resilience*), la législation[8] prévoit une information renforcée : **les personnes concernées, ou leurs représentants légaux, doivent être individuellement informés** (par la remise en main propre ou par envoi postal d'une note d'information).

Ce principe connaît toutefois des **dérogations** notamment lorsque les données ne sont pas directement collectées auprès de la personne et s'il est démontré que leur information individuelle exige des efforts disproportionnés notamment.

Dans tous les cas, le responsable de traitement doit mettre en place des **mesures de transparence complémentaires** comme par exemple la publication sur un site web de la note d'information relative au traitement, la mise en œuvre d'un portail de transparence, *etc.*

[6] La plateforme « Données & Design » du Laboratoire d'innovation numérique de la CNIL (le « LINC ») explique et illustre ce concept de transparence, avec des études de cas fictifs (ex : tableau de bord, parcours d'inscription, paramétrage de la géolocalisation …)
[7] Il s'agit des traitements de données de santé qui entrent dans le champ de la « section 3 » de la loi « informatique et libertés » modifiée (article 64 et suivants).
[8] La loi « informatique et libertés » modifiée (article 69)

La constitution de l'entrepôt par *Resilience* a nécessité qu'elle s'interroge sur la possible dérogation au principe de l'information individuelle des personnes notamment en raison du nombre de personnes concernées, de l'ancienneté des données ou de l'incertitude quant au statut vital de la personne.
Resilience a publié une note d'information (politique de confidentialité) distincte des conditions générales d'utilisation de ses produits et s'est interrogée sur la manière de rendre l'information la plus facilement compréhensible pour le public concerné (vulgarisation, utilisation d'images, présentation sous forme de tableaux, *etc.*).

- **Mettre en œuvre les droits.**

L'exercice des droits des personnes doit s'articuler avec le principe de minimisation des données. Ainsi les données d'identification ne doivent pas être collectées uniquement pour donner suite aux demandes d'exercice de droits des personnes.

Dans l'hypothèse d'un traitement de données pseudonymisées, le responsable doit démontrer qu'il n'est pas en mesure d'identifier la personne, l'informer de cette impossibilité et lui demander des informations complémentaires permettant de l'identifier.

Pour *Resilience,* les données sont pseudonymisées à la source. Cependant, il n'est pas exclu que *Resilience* soit en mesure de répondre à une demande d'exercice de droits, notamment si la personne lui fournit des informations complémentaires pour la réidentifier.

- **Vérifier la compatibilité de la réutilisation des données.**

Toute réutilisation de données personnelles suppose que les finalités ultérieures soient compatibles avec celles ayant justifié la collecte initiale. Pour cela, le responsable de traitement initial doit :

- avoir réalisé et documenté un **test de compatibilité des finalités**. Les finalités de recherche sont considérées comme compatibles par nature et ne nécessitent pas la réalisation de ce test ;
- donner son **accord écrit et spécifique** à cette réutilisation.

Resilience souhaite rassembler des données personnelles provenant de plusieurs traitements pour lesquels elle agit en qualité de responsable de traitement ou de sous-traitant. Dans certains cas, *Resilience* pourrait être destinataire des données, notamment pour les données issues de certains entrepôts hospitaliers. La mise à disposition des données par les établissements hospitaliers supposera qu'un test de compatibilité ait été réalisé et que ces derniers aient donné leur accord écrit et spécifique.

La création d'un entrepôt de données de santé nécessite le respect de certaines formalités. Afin de simplifier les démarches pour les responsables de ces bases de données sensibles, la CNIL a adopté en octobre 2021 un référentiel et, en complément, une « check-list » de conformité.

2. Comment articuler le concept d'intérêt public avec un traitement s'inscrivant dans un cadre commercial ?

Dans le domaine de la santé, il est fait référence à l'« *intérêt public* » dans différentes hypothèses, notamment concernant la base légale et la finalité.

La base légale du traitement

La base légale d'un traitement est ce qui autorise sa mise en œuvre, ou encore ce qui donne le droit à un organisme de traiter des données personnelles.

Le choix de la base légale la plus appropriée est essentiel pour **déterminer la formalité à réaliser** auprès de la CNIL à savoir l'autorisation ou la déclaration de conformité au référentiel « entrepôt ».

À noter : dans certains cas, aucune formalité n'est à réaliser auprès de la CNIL. C'est par exemple le cas lorsque le responsable de traitement recueille le consentement explicite de chaque personne concernée par l'enregistrement de leurs données dans l'entrepôt.

Resilience s'interrogeait sur la possibilité de fonder son traitement sur « *l'exécution d'une mission d'intérêt public* ». Toutefois, cette base légale est uniquement mobilisable par les **personnes morales de droit public ou les personnes morales de droit privé investies d'une mission de service public**. Ainsi, le projet de la jeune pousse devrait se fonder sur son(ses) **intérêt(s) légitime(s)**.

Le **consentement** des personnes a également été écarté dans la mesure où la société n'était pas en mesure de le recueillir pour chacune d'entre elles, notamment s'agissant de la réutilisation des données d'entrepôts constitués par des établissements de santé.

Resilience a fait le choix de solliciter une autorisation de la CNIL en prenant en compte les exigences techniques et juridiques du **référentiel « entrepôt »**.

Bien que ce référentiel soit uniquement applicable aux entrepôts de données de santé mis en œuvre par des personnes publiques, il fixe également la **doctrine de la CNIL en la matière.** Il est donc recommandé à tout responsable d'**en prendre connaissance et de s'attacher à en respecter le contenu.**

Le concept de « finalité d'intérêt public » dans le domaine de la santé

Tous les traitements de données personnelles dans le domaine de la santé comportent, par définition, des données de santé. La législation[9] prévoit que ces traitements doivent « *être mis en œuvre en considération de la **finalité d'intérêt public** qu'ils présentent* ».

Les **critères** permettant d'établir cette finalité d'intérêt public tiennent :

- d'une part, aux objectifs et bénéfices du traitement mis en œuvre ;
- d'autre part, aux modalités d'organisation envisagées (dans le cadre d'un entrepôt par exemple : l'objectif du traitement, les modalités de transparence, l'intégrité scientifique, la qualité des études, *etc.*).

Par principe, cette finalité d'intérêt public des traitements en santé n'est pas incompatible avec les intérêts – notamment commerciaux – d'une personne morale relevant du secteur privé.

Par exemple, une recherche médicale menée par une entreprise privée sur la COVID ou une application mobile d'éducation thérapeutique peuvent présenter un intérêt public.

À noter : la condition de « *finalité d'intérêt public* » ne correspond pas à la base légale du traitement (art. 6 du RGPD) mais à l'exigence prévue par les articles 44-3° et 66 de la loi « informatique et libertés » modifiée.

La CNIL a autorisé le projet d'entrepôt de *Resilience* estimant que le traitement poursuivait une finalité d'intérêt public (délibération n° 2022-049 du 21 avril 2022).

3. Quelle démarche pour utiliser les données d'un entrepôt à des fins de construction, de fonctionnement et d'amélioration continue d'un algorithme d'intelligence artificielle ?

Les données de santé contenues dans l'entrepôt constitué par *Resilience* (cf. question n° 1) ont vocation à être utilisées par un logiciel d'aide à la prise de décision médicale, fonctionnant notamment sur la base d'un algorithme d'IA.

[9] Article 66.I de la loi Informatique et Libertés

Dans le projet de *Resilience*, le développement, l'amélioration continue et le fonctionnement d'algorithmes d'IA impliquent de traiter des données de santé.

Dès lors que ce traitement de données de santé est mené dans **le domaine de la santé**, il peut être nécessaire de **réaliser de formalités préalables auprès de la CNIL**[10].

Les tableaux ci-dessous détaillent ces éventuelles formalités, en fonction des étapes de vie de l'algorithme et de la qualité du responsable de traitement.

La phase d'apprentissage

Cela correspond à la **phase de conception, développement et d'entraînement d'un système d'IA** utilisé par un logiciel (pré-marquage CE pour un logiciel d'aide à la décision médicale).

Qualification du traitement	Éventuelles formalités préalables auprès de la CNIL
Recherche, évaluation ou étude dans le domaine de la santé[11] **Pour déterminer si l'algorithme d'IA utilisé par le logiciel est au stade de développement**, plusieurs critères peuvent être pris en compte : caractère expérimental de l'algorithme, traitement mis en œuvre par des chercheurs, absence d'impact direct pour le patient (notamment pas de modification de sa prise en charge), logiciel pas ou pas encore mis sur le marché (par exemple pour les dispositifs médicaux, le produit n'a pas obtenu ou est en cours d'obtention du marquage CE).	**Formalités « recherche » :** • déclaration de conformité à une méthodologie de référence (MR-001, MR-003 ou MR-004) **ou** • demande d'autorisation auprès de la CNIL

La phase de production

Il s'agit de la phase de **déploiement opérationnel d'un système d'IA** en faisant fonctionner et en améliorant un algorithme d'IA utilisé par un logiciel d'aide à la décision médicale (post-marquage CE).

Qualification du traitement	Éventuelles formalités préalables auprès de la CNIL
Traitement mis en œuvre à des fins de « ***médecine préventive, des diagnostics médicaux, de l'administration de soins ou de traitement*** »[12] **par un professionnel de santé** soumis à une obligation de secret professionnel[13], et conformément à l'objectif pour lequel le logiciel (dispositif médical) a été créé (cela comprend son **amélioration continue**).	**Aucune formalité**[14] Toutefois, si le traitement n'est pas mis en œuvre sous la responsabilité du professionnel de santé prenant en charge le patient, le traitement est soumis à une **formalité « hors recherche »**[15] (aucune méthodologie de référence n'est applicable).

[10] Article 66 de la loi Informatique et Libertés
[11] Articles 72 et suivants de la loi Informatique et Libertés
[12] Article 44.1 de la loi Informatique et Libertés
[13] L'obligation de secret professionnel doit être distinguée de l'obligation de confidentialité.
[14] Articles 44.1 et 65.1 de la loi Informatique et Libertés
[15] Article 66 de la loi Informatique et Libertés

En pratique, deux conditions à respecter pour entrer dans cette qualification : 1. le logiciel est utilisé « **en vie courante** » à des fins de médecine préventive, de diagnostic, d'administration de soins ou de traitements ; 2. le logiciel est utilisé par un **professionnel de santé** prenant en charge la personne, responsable de traitement et **soumis au secret professionnel.** **Pour déterminer si le logiciel intégrant un algorithme d'IA est utilisé « en vie courante »**, plusieurs critères peuvent être pris en compte : utilisation par des professionnels de santé dans leur activité quotidienne, impact direct potentiel pour le patient, apprentissage intrinsèque de l'algorithme au fil de l'eau, utilisation du logiciel conforme aux spécifications de son marquage CE, etc.	

Pour approfondir

- Traitements de données de santé : comment faire la distinction entre un entrepôt et une recherche et quelles conséquences ?
- La CNIL adopte un référentiel sur les entrepôts de données de santé
- Entrepôts de données de santé : la CNIL publie une « check-list » de conformité à son référentiel
- Demandes d'autorisation en santé : la CNIL publie les critères à respecter
- Applications mobiles en santé et protection des données personnelles : Les questions à se poser

Les textes de référence

- Règlement général sur la protection des données (RGPD)
- Loi Informatique et Libertés
- Article L. 4001-3 du code de la santé publique (CSP)
- Délibération n° 2019-097 portant avis sur un projet de loi relatif à la bioéthique

Le projet « Magellan » : un outil contribuant à simplifier l'accès à des indicateurs de santé publique

Date de rédaction : juillet 2023

L'accompagnement de la société Clinityx, de fin juin 2021 à fin juin 2023, a porté sur la constitution de son **entrepôt de données de santé** dénommé « Magellan ».

Il s'agit du **premier entrepôt constitué uniquement à partir de certaines données du Système national des données de santé** (SNDS), qui rassemble des données de santé pseudonymisées issues des principales sources médico-administratives (remboursement de soins, activités hospitalières, etc.).

L'accompagnement de la CNIL s'est concentré sur la **déclinaison opérationnelle du principe de minimisation** mais aussi sur les **modalités de réutilisation des données personnelles** contenues dans cette base, et notamment la question des formalités applicables.

Les données de l'entrepôt ont vocation à alimenter un **outil de visualisation d'indicateurs agrégés** destiné à répondre à des questions de santé publique (outil « Magellan »). Les équipes de la CNIL ont travaillé avec Clinityx sur les **techniques d'anonymisation**.

Au terme de cet accompagnement, la CNIL a pu autoriser, en janvier 2022, la société Clinityx à constituer l'entrepôt « Magellan ».

La CNIL a également autorisé plusieurs projets de recherche réalisés à partir de cet entrepôt.

Les deux questions suivantes ont été traitées lors de l'accompagnement « bac à sable ».

1. Quel est le cadre juridique le plus adéquat pour le déploiement de l'outil « Magellan » ?

Les premiers échanges entre Clinityx et la CNIL ont permis de délimiter le projet et de distinguer les traitements de données de santé en découlant. De ce fait, il a été identifié que ce projet se déroulerait en deux temps :

1. **la constitution d'un entrepôt de données de santé** ;
2. lors de l'utilisation de l'outil « Magellan », **la mise en œuvre de traitements de données à des fins de recherches, d'études ou d'évaluations** dans le domaine de la santé.

Etape n° 1 : la constitution de l'entrepôt « Magellan »

Les points d'attention juridiques

En complément des dispositions du RGPD et de la loi Informatique et Libertés modifiée, le porteur de projet devait également respecter les **dispositions spécifiques applicables aux traitements de données du SNDS**.

Pour constituer cet entrepôt, le responsable de traitement a notamment dû :

- **définir la finalité**, c'est-à-dire l'objectif pour lequel les données sont collectées et utilisées au regard de la **plus-value scientifique et technique** apportée au traitement de données du SNDS.

Cet entrepôt a pour objet d'alimenter un outil « Magellan », destiné à calculer de façon automatisée des indicateurs prédéterminés de santé publique portant sur les populations de patients, le recours aux soins, et l'utilisation des produits de santé.

Plus précisément, la société Clinityx a déterminé les cas d'usage suivants :

- évaluer la représentativité d'un registre ou d'une cohorte ;
- évaluer l'accès aux produits de santé et à l'offre de soins pour la population ;
- évaluer l'apport d'une nouvelle technique, technologie, produit de santé ou pratique médicale sur la prise en charge d'une population ;
- évaluer l'impact des politiques de santé et de protection sociale sur la population ;
- réaliser des études de faisabilité dans le cadre d'une recherche impliquant ou n'impliquant pas la personne humaine.

L'utilisation des données de l'entrepôt pour un autre cas d'usage devra donc faire l'objet de formalités spécifiques auprès de la CNIL (par exemple pour réaliser des projets de recherche).

- **Minimiser les données**, c'est-à-dire collecter uniquement les données strictement nécessaires par rapport à l'objectif de cette collecte.

S'agissant du premier entrepôt constitué exclusivement à partir des données du SNDS, la création de celui-ci ne devait avoir ni pour objet, ni pour effet de créer au profit du responsable de traitement un accès permanent aux données du SNDS.

Le principe de minimisation a donc dû être décliné de façon opérationnelle au regard des cas d'usage associés à chaque objectif poursuivi par l'outil « Magellan ».

Pour le projet « Magellan », **le respect du principe de minimisation s'est matérialisé par l'identification des composantes du SNDS** (SNIIRAM, PMSI, etc.), nécessaires à l'alimentation de l'entrepôt **ainsi que de la profondeur historique des données**. Sur cette base, la liste des variables du SNDS nécessaires au fonctionnement de l'outil a été établie.

A l'issue de l'accompagnement, le porteur de projet a justifié scientifiquement tous ces éléments dans son dossier de demande d'autorisation. Pour cela, Clinityx a rempli une **expression de besoins** inspirée de celle transmise à la Caisse nationale de l'assurance maladie (CNAM) dans le cadre de la mise à disposition des données pour des projets de recherche.

En pratique, parmi les 240 variables issues du SNDS transmises par la CNAM, certaines d'entre elles (notamment celles destinées à rassembler les données d'un même patient, à calculer le coût d'une prise en charge ou à détecter des anomalies) sont conservées temporairement. **Seules 90 variables alimentent l'entrepôt de manière « pérenne ».**

- **Définir une durée de conservation** pour chaque catégorie de données et en fonction des finalités poursuivies.

Pour le projet « Magellan », la durée de conservation des variables est limitée à **cinq ans à compter de la mise à disposition des données par la CNAM**.
Cette conservation des données est réalisée en « fenêtre roulante ». Ce qui signifie que lors de chaque mise à jour des données, les données les plus anciennes sont anonymisées ou supprimées.

- **Informer les personnes concernées**[16]

En application des dispositions de l'article 14-5-b du RGPD, le responsable de traitement peut faire valoir une **exception à l'obligation d'information individuelle** pour la mise en œuvre d'un traitement comportant exclusivement des données issues du SNDS.

Dans cette hypothèse, il doit prendre des mesures appropriées pour protéger les droits et libertés ainsi que les intérêts légitimes des personnes concernées, y compris en rendant les informations

[16] Article 69 de la loi Informatique et Libertés

publiquement disponibles. Par exemple : la publication sur un site web de la note d'information relative au traitement, la mise en œuvre d'un portail de transparence, etc.

Pour le projet « Magellan », **un portail de transparence alimenté en temps réel** a été mis en ligne sur le site web de Clinityx. Il comporte une note d'information relative à la constitution de l'entrepôt et à son fonctionnement ainsi que toutes les notes d'information relatives à chaque requête réalisée.

Les mesures supplémentaires proposées

Mesures de transparence :

- tous les trois ans, un rapport sera transmis à la CNIL par la société Clinityx. Ce document portera sur le fonctionnement de l'entrepôt et les recherches réalisées à partir des données qu'il contient
- l'entrepôt sera inscrit au sein du répertoire public de la plateforme des données de santé (une fois celui-ci mis en ligne).

Gouvernance de l'entrepôt : des experts en matière de traitements de données du SNDS ont été intégrés à la comitologie de l'entrepôt afin d'aider le responsable de traitement à démontrer le respect des règles relatives à la protection des données tant durant le cycle de vie de l'entrepôt que de la réutilisation des données qu'il contient.

Les spécificités techniques et organisationnelles de l'entrepôt

Le dossier d'homologation

L'hébergement d'un entrepôt tel que « Magellan » en dehors de la plateforme sécurisée de la CNAM nécessite tout d'abord la création d'un **système fils du SNDS**.

Il doit respecter les **mesures de sécurité** mentionnées dans l'arrêté prévu par l'article L.1461-1 IV 3° du code de la santé publique relatif au **référentiel de sécurité du SNDS**[17].

Le respect de ce référentiel implique la constitution d'un **dossier d'homologation**, à transmettre à la CNIL. Ce dossier doit comporter les éléments suivants :

- un document démontrant la conformité, point par point, au référentiel SNDS ;
- un document d'homologation signé, établi par le(s) responsable(s) de traitement, attestant de sa (leur) connaissance du système d'information, des risques et des mesures de sécurité existantes et/ou prévues, et acceptant les risques résiduels ;
- le compte-rendu de la réunion d'homologation ;
- le plan d'actions détaillé pour la mise en place des mesures de sécurité prévues, actualisé récemment (notamment avec les acteurs, jalons, niveaux d'avancement).

L'architecture technique de la bulle sécurisée de la société Clinityx hébergeant l'entrepôt « Magellan » et utilisée pour l'hébergement de systèmes fils du SNDS a été analysée par la CNIL à diverses reprises, notamment dans le cadre de précédentes demandes d'autorisation. Une homologation de la bulle sécurisée a été réalisée par l'autorité d'homologation.

Cette bulle sécurisée est gérée par un hébergeur certifié pour l'hébergement de données de santé et est soumis exclusivement aux lois et juridictions de l'Union européenne.

[17] Il s'agit actuellement de l'arrêté du 22 mars 2017 relatif au référentiel de sécurité applicable au Système national des données de santé (en cours de mise à jour).

Les mesures techniques et organisationnelles

En cas de constitution d'un entrepôt comportant des données du SNDS, il est nécessaire d'appliquer les mesures techniques et organisationnelles prévues à la fois par le référentiel « entrepôt de données de santé » (délibération 2021-118 du 7 octobre 2021) et celles prévues par le référentiel de sécurité du SNDS.

Afin de de réduire les risques liés à l'utilisation de l'outil de requêtage, plusieurs mesures additionnelles ont été prévues avec Clinityx, et notamment :

- un accès à l'outil de requêtage et aux espaces projet limité aux personnes spécifiquement habilitées faisant partie du personnel interne de la société Clinityx, à l'exclusion de toute autre personne extérieure ;
- des mesures de cloisonnement des données, mises en œuvre *via* l'utilisation de solutions de conteneurisation logicielle, et permettant d'empêcher toute fusion de données ;
- des mesures techniques et organisationnelles permettant de différencier les accès aux données réalisés par les personnels accédant à la base de données de l'entrepôt « Magellan » contenant l'extraction des données du SNDS sur lequel repose l'outil de requêtage, de ceux réalisés par les personnels ne pouvant accéder qu'aux données minimisées contenues dans les espaces projet.

L'intégralité des mesures ont été formalisées dans :

- **une analyse d'impact relative à la protection des données** spécifique à l'entrepôt « Magellan » et aux différents espaces projet contenant les études liées à cet outil de production d'indicateurs ;
- **une analyse de risques sur la sécurité des systèmes d'information,** transmises dans le cadre de la demande d'autorisation relative à l'entrepôt.

Cet entrepôt n'étant pas conforme au référentiel « entrepôt de données de santé » de la CNIL, **Clinityx a déposé une demande d'autorisation durant son accompagnement « bac à sable ».**

Clinityx a pu s'appuyer sur les échanges avec la CNIL pour élaborer son dossier de demande d'autorisation.

Au regard de la complétude de son dossier, la CNIL a pu rapidement autoriser l'entrepôt pour une durée de 10 ans (délibération n° 2022-009 du 27 janvier 2022).

Etape n° 2 : la réutilisation des données issues de l'entrepôt « Magellan »

L'entrepôt « Magellan » alimente exclusivement l'outil « Magellan » qui ne précalcule pas les indicateurs, dans une **logique de protection des données par défaut**.

Par conséquent, pour fonctionner et générer des rapports statistiques **à la demande**, l'outil traite des données individuelles du SNDS.

Ces traitements de données constituent des recherches, études ou évaluations dans le domaine de la santé soumises aux dispositions de la loi « informatique et libertés » modifiée.

Ces traitements n'étant pas conformes aux méthodologies de référence s'agissant des traitements de données du SNDS, ils doivent être autorisés par la CNIL.

Plusieurs études ont ainsi pu être autorisés :

- une étude portant sur la chirurgie de l'incontinence urinaire d'effort (projet « Magellan Mesh »[18]) ;

[18] Délibération n° 2023-041 du 27 avril 2023 autorisant le Centre hospitalier universitaire de Poitiers et la société Clinityx à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité une étude portant sur le taux de réintervention pour complication et le taux de réintervention pour récidive après implantation de bandelettes sous-

- une étude portant sur la consommation en France de princeps et biosimilaires (projet « Anti-TNF alpha »[19]).

La CNIL a également délivré à Clinityx une **décision unique** (une autorisation pour réaliser plusieurs projets de recherche) pour mettre en œuvre, à certaines conditions[20], des études de faisabilité dans le domaine de la santé[21] à partir de l'entrepôt « Magellan ».

2. Quelles garanties pour aboutir à une anonymisation des indicateurs produits par l'outil ?

Pour pouvoir transmettre aux utilisateurs de l'outil des rapports statistiques générés après requêtage, Clinityx souhaitait s'assurer que ces rapports ne permettaient aucune ré-identification des personnes.

Les autorités de protection des données européennes définissent trois critères qui permettent de s'assurer qu'un jeu de données est véritablement anonyme[22] :

1. **l'individualisation** : il ne doit pas être possible d'isoler un individu dans le jeu de données ;
2. **la corrélation** : il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu ;
3. **l'inférence** : il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu.

Si ces trois critères ne peuvent être réunis, une analyse approfondie des risques de réidentification doit être menée par le responsable de traitement afin de démontrer que ces derniers, avec des moyens raisonnables, sont négligeables.

Les équipes de la CNIL ont principalement échangé avec Clinityx sur le processus détaillé permettant la production des indicateurs et leur présentation finale afin d'évaluer la nature des rapports produits par l'outil « Magellan ».

Clinityx a présenté l'interface de l'outil de requêtage et le modèle de la restitution fournie aux utilisateurs. La société a également fourni une **description des mécanismes et contrôles appliqués**.

Sur la base de l'avis n°05/2014 du G29 sur l'anonymisation, **la société a été invitée à affiner les opérations réalisées et à préciser le choix de certains paramètres.**

Enfin, il a été rappelé que les techniques de réidentification évoluent constamment et qu'une vigilance particulière devait être notamment portée aux requêtes proches réalisées par différents utilisateurs.

urétrales au cours d'une chirurgie de l'incontinence urinaire d'effort en distinguant les deux voies d'abord, intitulée « Magellan Mesh ».

[19] Délibération n° 2023-042 du 27 avril 2023 la société Clinityx a été autorisée à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité une étude portant sur la consommation en France de princeps et biosimilaires anti-TNF alpha, intitulée « Magellan Anti-TNF alpha ».

[20] Ces conditions sont prévues à l'article 66 IV de la loi Informatique et Libertés : « *La Commission nationale de l'informatique et des libertés peut, par décision unique, délivrer à un même demandeur une autorisation pour des traitements répondant à une même finalité, portant sur des catégories de données identiques et ayant des catégories de destinataires identiques.* »

[21] Délibération n° 2023-066 du 29 juin 2023 portant décision unique et autorisant la société « Clinityx » à mettre en œuvre des traitements automatisés ayant pour finalité la réalisation d'études de faisabilité dans le cadre d'une recherche impliquant ou n'impliquant pas la personne humaine, nécessitant un accès aux données de l'entrepôt « Magellan ».

[22] L'avis n°05/2014 sur les techniques d'anonymisation du G29

Le projet « Vertexa » : développer un jeu thérapeutique à destination des mineurs

Date de rédaction : octobre 2022

Les équipes de la CNIL ont accompagné la société *Vertexa* de fin juin 2021 à avril 2022 sur leur projet « VERTEXA (« *Virtual Reality Therapy Exposition in Anorexia* »).

Ce projet consiste à développer un **jeu thérapeutique en réalité virtuelle destiné à aider les patients atteints de troubles du comportement alimentaire** (TCA). La majorité des patients (personnes concernées) sont mineurs.

Le projet était en phase de développement au moment de l'accompagnement et des décisions étaient encore en cours sur le choix de certains prestataires (contenu des ateliers proposés aux patients, fabricant du casque de réalité virtuelle etc.). Initialement piloté au niveau du **Centre Hospitalier d'Arras,** ce projet est désormais porté par la start-up Vertexa, qui s'est structurée en société.

La CNIL a ainsi eu l'occasion de travailler avec un médecin chercheur Inserm, une chef de projet et la déléguée à la protection des données.
L'enjeu était de trouver **des solutions les plus adaptées en termes de protection des données** pour ce cas d'usage, impliquant le traitement de données de santé et des mineurs (population vulnérable).

Les trois questions suivantes ont été traitées lors de l'accompagnement « bac à sable ».

1. Information et consentement des patients (souvent mineurs) et de leurs parents : quelles modalités pour le cas d'usage ?

La première étape a été de définir **qui était responsable de traitement** lors de l'utilisation du jeu thérapeutique et de déterminer la base légale. La finalité de prise en charge médicale et la décision d'utiliser le logiciel appartient aux professionnels de santé/établissements de santé. Il a donc été admis que ceux-ci étaient responsables de traitement et la société Vertexa sous-traitante, en ce qu'elle héberge notamment les données générées par le jeu. C'est la base légale « intérêts légitimes » (article 6-1-e) qui a été considérée comme la plus appropriée.

Des efforts importants de transparence auprès des personnes concernées devaient être mis en œuvre pour tenir compte à la fois de la sensibilité du dispositif (pathologie, données de santé etc.) mais également de la population concernée principalement mineure, qui nécessitait d'adapter le langage et d'informer également les parents.

Il a été convenu d'adopter un processus dans lequel **le professionnel informe en premier lieu le patient et le parent à l'oral** (présentant de façon succincte et claire l'objectif et le fonctionnement du jeu thérapeutique ainsi que les données collectées et leur circuit). Pour assurer une traçabilité de l'information délivrée aux parents, le professionnel de santé documentera la délivrance de l'information et l'accord oral ou non opposition, via une case à cocher au sein de son interface. Il pourra compléter cette première information par la **remise d'un document écrit** auprès des titulaires de l'autorité parentale. Afin d'associer le mineur et créer un sentiment d'adhésion, celui-ci cochera également une case au sein du casque pour l'utilisation de ses données.

Pour compléter ce premier niveau d'information oral, **l'accessibilité de l'information RGPD est mise en œuvre via plusieurs canaux** : une vidéo diffusée au sein du casque avec les points essentiels, la note d'information complète disponible sur l'espace personnel du patient, sur le site Internet de Vertexa et au format papier sur demande auprès du professionnel de santé.

Pour aider le professionnel de santé dans son rôle de responsable de traitement, Vertexa prévoit de lui fournir **un guide** lui rappelant ses obligations RGPD et la manière dont l'information doit être délivrée aux patients.

Enfin, le porteur de projet a été encouragé à **anticiper la réutilisation des données** et à créer un portail de transparence destiné à répertorier les études qui seraient envisagées, en tant que responsable de traitement. Les documents d'information individuelle remis dans le cadre de la prise en charge renverraient vers ce portail de transparence. En anticipant de la sorte, le traitement de données pourrait alors entrer dans le cadre de la méthodologie de référence MR 004 ; ce qui permettra à Vertexa de bénéficier de formalités simplifiées.

2. Utilisation d'un casque de réalité virtuelle : comment sécuriser son utilisation ?

L'utilisation du casque de réalité virtuelle (VR) implique la collecte de données (création d'un profil, données d'identification, captation des mouvements, données comportementales, etc.). L'enjeu était donc de **créer une bulle sécurisée entre le patient et son professionnel de santé**, permettant de limiter au strict nécessaire la transmission au constructeur du casque des données générées lors de son utilisation, voire d'exclure toute transmission.

La première étape pour Vertexa était de **délimiter et définir le cas d'usage du casque** (à domicile ou chez le professionnel de santé, via un Wifi, etc.) **et d'identifier les flux de données** ainsi que les catégories de données collectées.

Il a aussi été recommandé à Vertexa de vérifier les fonctionnalités de base proposées dans le casque et ses conditions d'utilisation (authentification ou non auprès d'un service tiers, possibilité de prêt du casque etc.).

S'agissant plus particulièrement du développement de son application (qui sera sous-traité à un prestataire), Vertexa a été invité à être vigilant sur le choix des outils de développement et à conserver une certaine maîtrise sur les échanges de données engendrés par ceux-ci.

Une **demande de documentation auprès du fabricant du casque** devrait permettre de vérifier les éléments permettant d'offrir des garanties suffisantes pour la vie privée des personnes concernées. En outre, le circuit des données et les mesures de sécurité correspondantes devront être précisément décrits au sein de l'analyse d'impact relative à la protection des données (AIPD).

Il a été conclu que l'utilisation du casque ne pouvait être considérée comme sécurisée que si les conditions suivantes étaient remplies :

- l'absence de création de compte obligatoire auprès d'un tiers lors de l'utilisation du casque ;
- un logiciel « étanche » pour conserver les données en local eu égard à leur sensibilité ;
- une solution minimisant la collecte de données hors logiciel (OS, etc.) ;
- une solution sans transferts de données hors de l'Union européenne et qui privilégie le recours à un hébergeur soumis uniquement aux juridictions européennes.

3. Améliorer l'algorithme d'un logiciel médical : organiser l'utilisation et la conservation des données. Par où commencer ?

Si un organisme souhaite améliorer un logiciel médical (et particulièrement son algorithme), cela implique de réutiliser et parfois de centraliser les données issues de ce logiciel.

Avant de débuter, l'organisme doit répondre à plusieurs questions de « premier niveau » avant de poursuivre le déploiement ce traitement.

- **En quelle qualité agit l'éditeur de logiciel / Vertexa ?**
  S'il était sous-traitant de la finalité principale, il deviendra parfois responsable de traitement pour certaines finalités. Dès lors, certaines conditions devront être réunies et notamment une autorisation du responsable de traitement, conformément à la doctrine de la CNIL (sur ce point, voir la fiche pratique « Sous-traitants : la réutilisation de données confiées par un responsable de traitement »).

- **Quelles finalités ?**

- Lorsque la base de données est constituée pour permettre une amélioration continue de l'algorithme (apprentissage continu), le traitement entre dans le cadre de la prise en charge médicale par le professionnel de santé, qui reste responsable de traitement.

- En cas de développement d'un nouveau service ou logiciel, le traitement est assimilable à un traitement de recherche dans le domaine de la santé, et soumis aux dispositions spécifiques de la loi informatique et libertés et ses formalités. Dans cette hypothèse, Vertexa est considérée comme responsable de traitement.

- **Quelles catégories de données ?** Seules des données pertinentes au regard de la finalité envisagée doivent être collectées. Il conviendra d'opérer un tri au sein de l'ensemble des données recueillies dans le cadre du jeu thérapeutique Vertexa pour conserver uniquement les données strictement nécessaires pour cette réutilisation.

- **Comment informer les personnes ?** L'information des personnes peut être anticipée pour cette finalité dans le cadre de l'information initiale sur le dispositif et au sein du portail de transparence qui répertorie l'ensemble des études menées.
  Cette information doit permettre aux personnes concernées de s'opposer à cette réutilisation si elles le souhaitent.

Ces quatre points constituent un premier niveau d'analyse, qui peut être approfondi, notamment sur la question des formalités préalables auprès de la CNIL (sur ce point, voir les recommandations tirées du projet « *Résilience* » - question n° 3).