DMP du projet "Long term outcome of oesophagial atresia : transomics profiles in adolescence"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe : modèle structuré" fourni par Science Europe.

Renseignements sur le plan

Titre du plan	DMP du projet "Long term outcome of oesophagial atresia : transomics profiles in adolescence"
Livrable	D2.2
Version	Version initiale
Domaines de recherche (selon classification de l'OCDE)	
Langue	fra
Date de création	2022-10-13
Date de dernière modification	2022-12-05
Identifiant	PGD_TransEAsome_V1
Type d'identifiant	Identifiant local

Renseignements sur le projet

Titre du projet	Long term outcome of oesophagial atresia : transomics profiles in adolescence
Acronyme	TransEAsome
Sources de financement	• Agence Nationale de la Recherche : ANR-21-PMRB-0011
Date de début	2022-05-01
Date de fin	82028-05-01
Partenaires	• PROTEOMIQUE, REPONSE INFLAMMATOIRE ET SPECTROMETRIE DE MASSE (201119484K) • Bililles, Université de Lille () • Go@L, Université de Lille () • PedStart, INSERM Grand Ouest ()

Produits de recherche :

1. Données TransEAsome (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Gottrand Frédéric - 0000-0002-5290-0436	CHU de Lille	
Leroy Mélanie		• Personne contact pour les données • Responsable du plan de gestion de données
Non renseigné	Non renseigné	• Coordinateur du projet

DMP du projet "Long term outcome of oesophagial atresia : transomics profiles in adolescence"

1. Description des données et collecte ou réutilisation de données existantes

1.1 Description générale du produit de recherche

Nom	Données TransEAsome
Description	Les données cliniques dans le cadre de TransEAsome seront collectées par le logiciel Ennov Clinical via un cahier de recueil de données électronique, et concerneront à la fois des informations collectées dans le cadre du soin (histoire de la maladie, examen ou chirurgie réalisés, …) et des données spécifiquement collectées pour la recherche (questionnaires de qualité de vie). Des données omiques (transcriptomique, méthylomique et protéomique) seront générées suite à l'analyse d'échantillons biologiques (biopsies de l'œsophage et plasma) collectés conformément au protocole de l'étude. Les données omiques seront générées grâce aux machines Novasec et EVOSEP couplée à un orbitrap Q-Exactive (spectromètre de masse Thermo Scientific).
Type	Jeu de données
Workpackage	WP2, WP3,WP4 et WP5
Mots clés (texte libre)	

1.2 Est-ce que des données existantes seront réutilisées ?

1.3 Comment seront produites/collectées les nouvelles données ?

Description

Le projet TransEAsome a pour objectif d'inclure 300 patients et 150 témoins.

Les données cliniques ne seront collectées que pour les patients et seront intégrées à un electronic Clinical Report Form (eCRF) développé par l'unité Statistique, Évaluation Économique, Data-management (SEED) du CHU de Lille via le système Ennov Clinical. Le contenu de l'eCRF pourra être exporté au format Excel, avec 450 lignes (une par participants) et autant de colonnes que de données de santé.

Les données omiques brutes seront aux formats fastq (brutes), bam (alignés sur le génome) et/ou csv.

Les données brutes en sortie du spectromètre de masse seront sous un format Xcalibur Raw File (.raw) et d'une taille d'environ 1,5 Go par échantillon. Ces fichiers bruts seront traités par le logiciel MaxQuant qui produira des fichiers résultats sous format texte (.txt) d'une taille d'environ 2 To. Ces fichiers résultats seront utilisés par Perseus (.SPS), permettant de générer des fichiers processés de type tableur (.csv) sous forme d'une matrice contenant différente colonne (valeur d'intensité, liste de gêne, protéine, etc). Des comptes rendus sous format power point, word ou pdf seront également générés contenant des images et des graphiques (10 à 50 Mo).

Les rapports d'analyse multi-omiques ainsi que d'éventuelles analyses complémentaires en single-omique seront principalement produits sous forme de document R markdown (.rmd), permettant de générer à la demande des fichiers aux formats usuels .pdf ou .html intégrant textes et graphiques (~10 à 50 Mo par rapport pdf/html; <~10 rapports produits). Certains de ces rapports pourront être accompagnés de fichiers tableurs additionnels (.csv/.tsv) contenant des valeurs numériques et/ou textuelles (liste de gènes, listes de voies métaboliques, statistiques sur les données) afin de faciliter l'automatisation des traitements (généralement <50Mo pour chacun de ces fichiers additionnels). Ponctuellement des rapports plus spécifiques pourront être générées sous forme de documents texte (.odt ou .doc) tels que des rapports de réunion (<10 Mo).

Le projet prévoit l'intégration dans l'environnement France Cohortes de données cliniques (environ 50 Mo de données issues de questionnaires) pour la base clinique à partir de 2026, et de données omiques brutes (données protéomiques (1,5 To), données de transcriptomique et de méthylation (6 To)) à partir de 2027.

---

2. Documentation et qualité des données

2.1 Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?

2.2 Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?

Description

Les données de santé seront monitorées par un.e attaché.e de recherche clinique du CHU de Lille selon le plan de monitoring du protocole. En complément des contrôles de cohérences seront mise en place au sein de l'eCRF par le data manager (validité des formats de données, fourchettes de valeurs, cohérence des dates, …), et complétés par des contrôles spécifiques demandés par le promoteur.

Tous les tests seront décrits dans le plan de validation de données.

Trimestriellement /Mensuellement, les tests programmés seront exécutés sur les pages monitorées de chaque eCRF. Le data manager éditera ces queries par patient et les rendra accessibles aux centres.

Le contrôle de la qualité des données protéomique de spectrométrie de masse sera réalisé par l'analyse de plusieurs échantillons control au sein de la séquence complète. Ces analyses seront aussi contrôlées par des analyses non supervisées de type PCA et distribution normale.

Le contrôle de la qualité des données omiques générées sera également assuré par des analyses statistiques standards de type non-supervisé (contrôle des distributions, ACP,…) permettant de détecter de potentiels biais techniques liés à la génération de ces données.

---

3. Exigences légales et éthiques, code de conduite

3.1 Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?

Description

Les données de santé seront stockées sur un serveur sécurisé du CHU de Lille (certifié ISO 27 000). La Direction des Ressources Numériques du CHU de Lille assure l'hébergement des données cliniques sur un serveur sécurisé, ainsi que leur sauvegarde quotidienne. Elles seront stockées pendant 2 ans après la publication des résultats de l'étude, conformément à la réglementation en vigueur, et seront archivées pendant 15 ans conformément aux Bonnes Pratiques Cliniques.

Les données omiques et les métadonnées associées seront stockées dès leur génération et à minima jusqu'à xx années après la fin du projet (à quel , sur chaque PF ? Uniformisé ou non ? Quel moyen de partage de ces données ? Qui supervise l'ensemble ?)

Concernant les analyses multi-omiques, les codes sources des analyses et les rapports présentant les résultats seront stockés dès leur génération et jusqu'à XX années après la fin du projet sur le GitLab de l'université de Lille ainsi que sur un serveur interne à l'UAR 2014- US 41 PLBS géré par la plateforme SINBIOS de cette même unité.

De manière temporaire et selon les besoins des analyses, les données cliniques et données omiques pourront être stockées sur les serveurs de l'UAR PLBS, le cloud de la plateforme bilille hébergé par le mésocentre de l'Université de Lille, ainsi que sur les disques durs cryptés des ingénieurs de la plateforme bilille. Ce stockage temporaire pourra s'étendre jusqu'à la durée du projet, au-delà le stockage de ces données ne pourra excéder 6 mois dès lors que les analyses additionnelles auront été réalisées.

En fin de projet, les données seront stockées sur la plateforme France Cohortes. La plateforme de France Cohortes est physiquement hébergée au Datacenter de niveau Tiers III du CINES à Montpellier dans deux salles séparées hébergeant l'infrastructure de production du système France Cohortes : SM1 et SM3. Les schémas de mise en rack et de la connectivité entre les Différents Composants de la plateforme sont disponibles dans le DAT du SI France Cohortes. Le site de PRA (Plan de Reprise d'Activité) quant à lui se situe à Lognes dans le Datacenter EUCLYDE DC6 certifié HDS.

Une fois que les données sont mises en qualité, épurées et enrichies, elles seront régulièrement versionnées et historisées dans l'entrepôt de données.

Les sauvegardes sont effectuées à minima sur une base quotidienne en horaires non ouvrés. Un document regroupant la politique de sauvegarde contient les fréquences et les durées de rétention par type de données.

La confidentialité et l'intégrité des données sauvegardées sont assurées par les différents mécanismes suivants :

- Sécurisation des données de santé à caractère personnel (chiffrement AES256) ;
- Contrôle des accès logiques et physiques ;
- Chiffrement du transport ou liaison privées en cas d'externalisation sur un site secondaire, et le chiffrement des supports amovibles en cas d'externalisation des sauvegardes ;
- Opérations de contrôle et gestion des incidents.

Les rapports produits par les logiciels de sauvegarde sont vérifiés quotidiennement et conservés. L'architecture de sauvegarde cible prévoit une externalisation des données de sauvegarde de France Cohortes sur un Datacenter distant de l'Inserm. Cette externalisation se fait au travers d'une liaison réseau chiffrée. Il y a systématiquement deux types de sauvegarde, les sauvegardes systèmes et les sauvegardes applicatives (contenant les données).

Des tests mensuels de restauration aléatoires de données sont réalisés dans le but de garantir la validité des sauvegardes. Les sauvegardes et restaurations sont actuellement effectuées par un outil de sauvegarde conforme aux standards du marché. La durabilité des données est garantie pendant une durée définie conformément aux autorisations réglementaires obtenues.

3.2 Quelles sont les contraintes juridiques (sensibilité des données autres qu'à caractère personnel, confidentialité, ...) à prendre en compte pour le partage et le stockage des données ?

Description	Les données de santé pseudoanonymisées (utilisation d'un code d'inclusion) ne seront accessibles qu'aux personnes avec un accès à l'eCRF (identification personnelle et étendue des droits d'accès spécifiques). Les accès sont gérés par l'équipe de datamanagement du CHU de Lille, et tracés via des formulaires de demande d'accès et des audits trails de connexion. Les accès ne pourront être fournis qu'aux membres des équipes partenaires. La connexion à l'eCRF est assurée par un code étude, un identifiant individuel ainsi qu'un mot de passe personnalisé. Seuls les data managers et le promoteur ont accès à toutes les données de l'étude. Les investigateurs, de leur côté, ont un accès aux dossiers patients limité au(x) centre(s) qui les concerne(nt). Seuls les data managers ont accès à la base de données, et avec une double authentification. Conformément au règlement européen de protection des données, une analyse d'impact sera réalisée pour les données de santé et sera conservée par le délégué à la protection des données du CHU de Lille. Données omiques sur les serveurs de l'université ? Faire préciser les caractéristiques du stockage proposé par le mésocentre d'U.Lille (cloud bilille, stockage interne UAR PLBS)

3.3 Quels sont les aspects éthiques à prendre en compte lors de la collecte des données ?

Description	Dans le cadre des données de santé collectées dans le cadre du protocole, un consentement éclairé sera signé par les parents des patients souhaitant participer à l'étude. Le modèle du formulaire de consentement sera validé par un Comité de Protection des Personnes (CPP) et indiquera la procédure pour exercer leurs droits prévus par le RGPD Le projet TransEAsome sera soumis à un Comité de Protection des Personnes et respectera les Bonnes Pratiques Cliniques.

4. Traitement et analyse des données

4.1 Comment et avec quels moyens seront traitées les données ?

Question sans réponse.

5. Stockage et sauvegarde des données pendant le processus de recherche

5.1 Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

Besoins de stockage

Les données cliniques et omiques brutes seront archivées sur la plateforme France Cohortes et rendues disponibles après la levée de l'embrago. Les données seront réutilisables une fois que les responsables de traitement auront finalisé et valorisé leurs travaux. Seule une sélection des données choisies par l'équipe porteuse du projet pourra être partagée.

Un process bien défini permettra l'accès à ces données. Un système de gestion des accès aux données sera mis en place par l'intermédiaire de France Cohortes, dont le processus sera :

– Dépôt d'une demande de justifications éthique et scientifique ;
– Évaluation de la demande par comité scientifique de la cohorte ;
– Si réponse positive, démarche réglementaire pour obtention des droits d'accès ;
– Création de bulles sécurisées pour l'équipe demandeuse ;
– Extraction mise à disposition des données demandées pour un temps limité défini par le projet ;
– Valorisation et communication autour de ces données.

Les données doivent être conservées pendant une durée qui n'excède pas la durée nécessaire aux finalités pour lesquelles elles sont collectées et traitées (30 ans au maximum pour les données de santé), sauf en cas de demande expresse de suppression formulée par le patient. Pour les traces (logs), la conservation ne doit pas dépasser 4 ans.

Les gestionnaires du système d'information SI France Cohortes doivent être en mesure de restituer l'ensemble des données sauvegardées et archivées. Par ailleurs, de nombreuses actions malveillantes sont réalisées sur les données hors-ligne, il convient donc de s'assurer qu'elles bénéficient d'un niveau de sécurité optimal.

Les données seront mises à disposition sur des environnements sécurisés sur la plateforme France Cohorte, certifiée HDS. La sauvegarde des environnements VxRail est assuré par CommVault en version 11.20. La sauvegarde principale est effectuée sur une baie de stockage Netapp FAS8020 puis répliqué sur un environnement Scality pour l'archivage des sauvegardes.

Le type de sauvegarde sera est du [Network Block Device]. La sauvegarde est assurée à travers le réseau. L'ensemble des composants virtuels pour la sauvegarde se trouve sous le répertoire SVG de l'environnement VxRail. Une sauvegarde et une copie auxiliaire est prévue :

- Sauvegarde principale (tous les DC):

C'est une sauvegarde de type « Incremental Forever » avec :

-Une sauvegarde incrémentale lancée tous les jours avec une rétention de 30 jours.

-Une sauvegarde de type « Synthetic Full » lancée automatiquement tous les 15 jours avec une rétention de 30 jours.

Une sauvegarde complète Synthétique est un job administratif qui ne sollicite pas le client. Elle tourne en journée et permet de fermer les cycles de sauvegarde. Cette sauvegarde autorise une restauration sur 44 jours.

- Copie auxiliaire vers S3 Scality:

La première Full du mois sera copiée sur le stockage S3 avec une rétention d'un an. Une sauvegarde annuelle pourra également être définie.

Des restaurations de fichiers de machines virtuelles seront également possibles selon un process détaillé dans le DAT du SI de France Cohortes.

A long terme une méthode d'archivage sera proposée.

Volume estimé des données

0

---

6. Partage des données et conservation à long terme

6.1 Comment les données seront-elles partagées ?

Question sans réponse.

6.2 Comment les données seront-elles conservées à long terme ?

Justification

Les données seront mises à disposition sur des environnements sécurisés sur la plateforme France Cohorte, certifiée HDS. La sauvegarde des environnements VxRail est assuré par CommVault en version 11.20. La sauvegarde principale est effectuée sur une baie de stockage Netapp FAS8020 puis répliqué sur un environnement Scality pour l'archivage des sauvegardes.
Le type de sauvegarde sera est du [Network Block Device]. La sauvegarde est assurée à travers le réseau. L'ensemble des composants virtuels pour la sauvegarde se trouve sous le répertoire SVG de l'environnement VxRail. Une sauvegarde et une copie auxiliaire est prévue :

- Sauvegarde principale (tous les DC):

C'est une sauvegarde de type « Incremental Forever » avec :
-Une sauvegarde incrémentale lancée tous les jours avec une rétention de 30 jours.
-Une sauvegarde de type « Synthetic Full » lancée automatiquement tous les 15 jours avec une rétention de 30 jours.
Une sauvegarde complète Synthétique est un job administratif qui ne sollicite pas le client. Elle tourne en journée et permet de fermer les cycles de sauvegarde. Cette sauvegarde autorise une restauration sur 44 jours.

- Copie auxiliaire vers S3 Scality:

La première Full du mois sera copiée sur le stockage S3 avec une rétention d'un an. Une sauvegarde annuelle pourra également être définie.
Des restaurations de fichiers de machines virtuelles seront également possibles selon un process détaillé dans le DAT du SI de France Cohortes.

A long terme une méthode d'archivage sera proposée.

Volume estimé des données 0

Archive :