DMP du projet "PrediMAP : Développement et évaluation clinique d'un dispositif médical innovant pour prédire l’accouchement prématuré - De la recherche fondamentale aux urgences obstétricales"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "PrediMAP : Développement et évaluation clinique d'un dispositif médical innovant pour prédire l’accouchement prématuré - De la recherche fondamentale aux urgences obstétricales"
Livrable	DMP n°1
Version	Version initiale
Domaines de recherche (selon classification de l'OCDE)	Clinical medicine, Medical engineering, Health sciences, Health biotechnology
Langue	fra
Date de création	2022-09-21
Date de dernière modification	2022-10-05

Renseignements sur le projet

Titre du projet	PrediMAP : Développement et évaluation clinique d'un dispositif médical innovant pour prédire l'accouchement prématuré - De la recherche fondamentale aux urgences obstétricales
Acronyme	PrediMAP
Résumé	PrediMAP est une étude de performance diagnostique d'un DM-DIV non comparative, non contrôlée. Son objectif principal est de développer et de valider cliniquement la performance prédictive d'un algorithme et du DM-DIV PrediMAP pour prédire l'accouchement dans les 7 jours dans la population cible des femmes consultant aux urgences obstétricales pour une menace d'accouchement prématurée (MAP). C'est une étude qui se déroule en 3 étapes (3 cohortes). Le nombre total de patientes à inclure est de 3600 pour une durée totale de l'étude de 4 ans et demi
Date de début	2022-09-01
Date de fin	2027-08-31
Partenaires	• DMU APHP.Centre : Femme-Mère-Enfant (201923403K) • BforCure () • Université Paris Cité () • DMU APHP.Nord : GYNECOLOGIE PERINATALOGIE PARIS NORD (201923380K) • DMU APHP.Sorbonne : Femmes Mères Enfants (Obstétrique Reproduction Infertilité Gynécologie Enfants) (201923350C) • Institut national de la santé et de la recherche médicale ()

Produits de recherche :

1. Dispositif Médical de Diagnostic In Vitro PrediMAP et algorithme prédictifs de la menace d'accouchement prématuré (Objet physique)

Contributeurs

Nom	Affiliation	Rôles
GOFFINET François		• Coordinateur du projet • Personne contact pour les données
MESBAHI-IHADJADENE Karima		• Responsable du plan de gestion de données

DMP du projet "PrediMAP : Développement et évaluation clinique d'un dispositif médical innovant pour prédire l'accouchement prématuré - De la recherche fondamentale aux urgences obstétricales"

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

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Le recueil de toutes les données de la recherche clinique sera réalisé dans le logiciel sécurisé Cleanweb à partir des données de plusieurs sources.
Les données cliniques sont saisies directement par l'utilisateur (Investigateur, TEC…) dans l'e-CRF CleanWEB.Les données « brutes » sont ensuite exportées de l'e-CRF par le data manager de l'étude.
Des données biologiques seront produites lors de la recherche au laboratoire de l'Institut Cochin par les membres du groupe de Céline Méhats (Postdoctorante, Ingénieurs d'étude) sous la responsabilité de C Méhats. Toutes les expériences et leurs résultats seront rapportés sur le cahier de laboratoire électronique (CLE Inserm).

1) Données inédites de quantité de molécules générées par dosage ou par immuno-PCR dans les sécrétions cervico-vaginales générées par dosage ELISA et analyse spectrophotométrique, les données brutes seront importées du spectrophotomètre et stockées au laboratoire sous forme de fichiers excel.

2) images de tissus humains inédites qui seront obtenues après immunofluorescence multiplex pour 30 femmes. Ces données seront recueillies par un scanner de lames (Lamina slide scanner" haut débit, Akoya Perkin Elmer). Les lames virtuelles générées par le scanner seront importées et stockées dans la CID (Cochin Image Database) sur un site dédié et dont l'accès est protégé.

3) recueil de données génétiques inédites (séquences d'ARN) par transcriptomique spatiale et séquençage à haut débit (NextSeq™ 500 Illumina) pour 30 femmes. Les données brutes seront traitées avec Space Ranger pour réaliser les alignements et les comptages des UMI (*Unique Molecule Identifier*) et la génération des matrices associant coordonnées spatiales et données traitées.
Les données cliniques et de mesures des biomarqueurs dans les sécrétions cervico-vaginales de la cohorte Inspire,seront réutilisées pour les analyses et la réalisation de l'algorithme.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

- Les données recueillies et leur format sont préalablement définis et validés dans un document « liste des variables »
- Les données cliniques et biologiques directement saisies sur l'eCRF seront stockées dans la base de donnée CleanWEB. Ces données seront ensuite extraites dans plusieurs fichiers .csv afin de les exploiter pour les besoins de l'analyses statistique.
- Images de microscopie aux formats .tiff et .PNG
- Matrices au format .txt, Tableaux au format Excel pour les besoins des analyses bioinformatiques et statistiques et documents texte en .docx.

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

- MDH (Manuel de Data Handling, dictionnaire des données) : structure de l'export final avec le format des données
- CRF annoté généré depuis CleanWEB
- Métadonnées : données de séquences et leur quantité, coordonnées spatiales et phénotype de la patiente (type d'accouchement, terme d'accouchement, etc.).

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Pour les données cliniques et biologiques :

1. Monitoring : l'étude a été classée à risque minimal (risque A). L'attaché de recherche clinique n'effectuera pas de contrôle des données saisies sur l'eCRF par rapport aux données sources.
2. Data management : 3 types de contrôles sont prévus dans le plan de data management :
    - Queries automatiques CleanWeb exécutés périodiquement (lot de queries sur l'eCRF)
    - Contrôle lors du saisie exécuté en temps réel dès la saisis
    - Contrôle à posteriori par le statisticien

A l'issue de chaque lot de queries, un rapport de data management est rédigé par le data manager.Ce rapport contient différents indicateurs sur le taux de saisie et l'évolution du statut des queries.

Pour les données biologiques

Les analyses bioinformatiques et statistiques seront réalisées régulièrement. Les fichiers seront analysés par au moins deux investigateurs de façon indépendante.Des rapports réguliers seront produits qui contiendront différents indicateurs de traçabilité et degré de confiance des données et résultats obtenus.

Du côté de BForCure, les algorithmes seront développés par le pôle Intelligence Artificielle (IA), et la mise à disposition sous forme de produit logiciel sera assuré par le pôle Software. Des analyses de données plus poussées pourront être faites par le pôle IA.

3. Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Les eCRF Cleanweb sont stockés dans un serveur de l'APHP certifié hébergeur de données de santé.

Les données biologiques seront stockées dans trois disques durs externes conservés dans des endroits distincts à accès contrôlé et dans un serveur dédié à la sauvegarde de l'Institut Cochin.Les données génétiques seront stockées sous forme de fichier .txt (Bam et fastq) dans trois disques durs conservés dans des lieux différents et protégés par un code d'accès et pourront être transférées en fin d'étude à une organisation européenne (*European Genome Archive*) pour archivage dans l'intérêt public, à des fins de recherche scientifique, sous réserve que les patientes ne s'y soient pas opposées après en avoir été informées.

L'hébergement de GAIA est assuré par un herbergeur certifié « hébergeur de données de santé », avec une localisation des données assurées en Union Européenne, à travers un entrepôt de données de type SQL. Une sauvegarde est effectuée toutes les nuits, avec un format agnostique du moteur de base de données (JSON). La méthode utilisée pour le stockage permet de revenir en arrière à tout moment et de lister les actions effectuées.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Données cliniques :

Dans l'interface CleanWeb standard, les données sensibles peuvent être recueillies au niveau de variables spéciales de type "DONNÉE PERSONNELLE" ; elles sont alors :

- stockées de manière cryptée dans une base séparée de la base étude
- accessibles de manière non-cryptée seulement sous certaines conditions (pour les utilisateurs du centre auquel est rattaché le patient)
- non-exportables
- non-utilisables comme élément de la référence patient.

Les données personnelles de type "Nom" et "Prénom" seront affichées sur la liste des patients par défaut et le champ de recherche filtrera sur la référence ainsi que sur les colonnes de type "Nom" et "Prénom" si l'utilisateur peut les visualiser.

Dans l'interface CleanWeb ePRO, les données de contact (email ou téléphone) du patient :

- ne font pas partie du cahier d'observation, mais ont un accès assujetti à un droit spécifique
- sont stockées dans la base étude dans deux colonnes séparées
- sont cryptées en base, et lors de la connexion

- sont non-exportables, non-imprimables, visibles et éditables uniquement au niveau du calendrier patient
- sont supprimées de manière définitive à la fin de l'étude, lorsque le statut de l'étude est coché "Terminée"
- peuvent être effacées par patient au fur et à mesure de l'avancement de l'étude si besoin.

Lorsqu'un utilisateur est habilité à saisir des données dans l'eCRF (investigateur ou TEC ayant remis son CV, son attestation BPC < 5 ans, ayant été formé (training log complété) et étant inscrit sur le Formulaire de Délégation de Fonction), le chef de projet et/ou son Assistant (ACP) et/ou l'ARC de l'étude demande au data manager de lui donner un accès (selon les paramètres définis dans la liste des « intervenants et comptes eCRF » et validés par le chef de projet).

L'utilisateur reçoit un lien vers un site sécurisé et la dernière version du guide utilisateur. Il créé lui-même son mot de passe à partir de ce lien, en respectant la norme préconisée par la CNIL (8 caractères au moins, différent de l'identifiant ainsi que des nom et prénom, incluant au moins une majuscule, au moins une minuscule, au moins un chiffre et au moins un symbole parmi les suivants : ! ? $ % ( ) _ + - [ ] { } : ; @ § # < > /.)

Après connexion à l'interface, si l'utilisateur est inactif sur son compte pendant plus de 15 min, une déconnexion automatique est programmée. Il devra alors se reconnecter en utilisant ses codes d'accès.

Par ailleurs, trois saisies répétées d'identifiants erronés provoquent le blocage du compte de l'utilisateur pendant 20 minutes, afin d'éviter un accès non autorisé à l'interface. Passé ce délai, l'utilisateur pourra de nouveau tenter de se connecter.

L'utilisateur communique avec le serveur via une connexion internet sécurisée utilisant les protocoles standards de l'internet : HTTPS avec cryptage SSL.Un navigateur Internet quelconque fournit l'interface utilisateur ; l'utilisateur doit s'assurer que la connexion Internet fonctionne.Avant de pouvoir se connecter, l'utilisateur disposera de ses codes d'accès : un identifiant (login) et un mot de passe.

Les codes d'accès seront demandés à chaque connexion : ils permettent à l'utilisateur de s'authentifier sur le système et d'accéder alors aux seules données et actions le concernant.

Les serveurs de l'APHP ont une sauvegarde journalière.

Une fois la base de données de l'étude constituée une copie de la base est sauvegardée sur le serveur.

L'extraction finale et intégrale des données est faite par le data manager à partir du module d'extraction de CleanWEB.

Les données brutes sont structurées avec les outils de data management selon le format prévu dans le MDH de l'étude.

La transmission des données est faite par l'intermédiaire d'un lien sécurisé et valable 30 jours qui permet au destinataire de télécharger les données à partir du serveur dispose.aphp.fr

Le lien et les codes d'accès sont envoyés dans des mails séparés.

Les données sont elles-mêmes stockées dans une archive zip qui est aussi protégée par un mot de passe.

Les transmissions de données sont tracées dans un tableau (DTYP URC) intitulé « suivi-transmission-base ».

Données biologiques

Pour les données biologiques, les échantillons sont pseudo-anonymisés avant leur transfert au laboratoire par le personnel de l'URC et référencés sous un code au laboratoire (date du recueil, numéro d'arrivée). Un registre listant les patientes incluses dans le protocole répertorie par leur code le nombre et la nature des échantillons extraits des prélèvements. Le registre sera conservé dans un placard fermé à clef et à accès restreint. Les images et les dosages de protéines ne constituent pas de données sensibles.

Les données de séquences ne seront accessibles qu'à des personnes habilitées. A la fin de l'étude, les métadonnées (données de séquences et leur quantité, coordonnées spatiales et phénotype de la patiente (type d'accouchement, terme d'accouchement, etc.) seront transmises et archivées via le site européen EGA à des fins de recherche scientifique dans l'intérêt public. L'EGA fournit une solution d'archivage entièrement gratuite, sécurisée et permanente pour le partage de données dans le monde entier. Les déposants conservent l'entière propriété des données et peuvent soumettre les données par étapes et contrôler les autorisations d'accès aux données une fois soumises. Chaque étude se voit attribuer un numéro d'accès stable et unique auquel il pourra être fait référence dans de futures publications. L'EGA accepte les données anonymisées avec un plan approuvé par un comité d'accès aux données (Data Access Consortium : DAC), qui est responsable de toutes les décisions d'accès aux données. Le DAC, chargé de prendre les décisions d'accès aux données soumises sera composé du chercheur en charge de recueil des données génétiques.

La configuration de l'EGA consiste en une installation informatique sécurisée pour le traitement des données et une configuration EBI partagée pour la soumission et la distribution des données via des demandes de données effectuées via le site Web de l'EGA. Toutes les données distribuées sont cryptées et ne sont accessibles qu'à l'aide d'une clé de cryptage, qui est distribuée aux utilisateurs par la poste ou par courrier.

Concernant GAIA, il n'y a besoin d'aucune donnée personnelle, seul un identifiant d'échantillon unique est nécessaire, mais sans connexion directe à un système externe permettant d'identifier le patient. Aucun outil n'est à installer, puisqu'un navigateur web est à disposition. Si nécessaire, un SDK python peut être fourni pour mieux programmer des analyses de données.

---

4. Exigences légales et éthiques, codes de conduite

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Le fichier informatique utilisé pour cette Investigation Clinique est mis en œuvre conformément à la règlementation française (loi Informatique et Libertés modifiée) et européenne (Règlement Général sur la Protection des Données –RGPD).

Conformément aux dispositions de l'article 62, paragraphe 4f du Règlement Européen 2017/745, aucune Investigation Clinique ne peut

être pratiquée sur une personne sans son consentement libre et éclairé, recueilli par écrit après que lui a été délivrée l'information prévue à l'article 63, paragraphe 2 du Règlement Européen 2017/745. Le consentement libre, éclairé et écrit de la personne est recueilli par l'investigateur principal ou par un collaborateur déclaré et formé à l'investigation clinique avant l'inclusion de la personne dans la recherche.

Les sujets seront identifiés dans Cleanweb de la façon suivante :

n° du centre (3 positions numériques) – le numéro de la cohorte (deux positions numériques) - n° d'ordre de sélection de la personne dans la cohorte (4 positions numériques) - initiales nom - initiale prénom.

Centre : |__|__|__| Patient : |__|__|-|__|__|__|__|- |__|- |__|

Cette référence est unique et sera conservée pour toute la durée de l'Investigation Clinique.

Lors des transmissions de données sur la plateforme sécurisée de BForCure, afin de garantir la pseudo-anonymisation des données le numéro d'identification d'export comportera les seuls éléments suivants : n° centre (3 positions numériques) – le numéro de la cohorte (deux positions numériques) - n°ordre de sélection de la personne dans la cohorte (4 positions numériques). Les initiales ne seront pas transmises.

Centre : |__|__|__| Patient : |__|__|-|__|__|__|__|

A l'issue de l'Investigation Clinique, les échantillons pourront être utilisés pour des analyses ultérieures non prévues dans le protocole pouvant se révéler intéressantes dans le cadre de la grossesse et de ses complications, en fonction de l'évolution des connaissances scientifiques, sous réserve que la patiente ne s'y soit pas opposée, après en avoir été informée, comme indiqué dans le formulaire d'information/consentement. Les patientes seront informées que les données génétiques feront l'objet d'un traitement informatique et pourront être transférées à une organisation européenne pour archivage dans l'intérêt public, à des fins de recherche scientifique, sous réserve qu'elles ne s'y opposent pas, comme indiqué dans le formulaire de consentement conformément aux articles 13 et 14 du RGPD.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Chaque partenaire est propriétaire de ses données (données fournies ou données produites). Les autres dispositions de propriété intellectuelle concernant les données produites dans le cadre du projet sont encore en cours de discussion pour l'établissement du contrat de consortium.

L'AP-HP est propriétaire des données cliniques et aucune utilisation ou transmission à un tiers ne peut être effectuée sans son accord préalable.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

L'AP-HP en tant que promoteur obtient pour l'Investigation Clinique sur un dispositif médical, préalablement à sa mise en œuvre l'avis favorable du CPP concerné, dans le cadre de ses compétences et conformément aux dispositions législatives et réglementaires en vigueur.

---

5. Partage des données et conservation à long terme

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Les données cliniques seront conservées en base active jusqu'à 2 ans après la dernière publication de la recherche.

Les données biologiques seront conservées sur le CLE jusqu'à 10 ans après la dernière publication de recherche. Les métadonnées (données de séquences et leur quantité, coordonnées spatiales et phénotype de la patiente (type d'accouchement, terme d'accouchement, etc.) être archivées sur le site d'archivage EGA. Un DAC contrôlera l'accès aux données anonymisées à des tiers pour des recherches scientifiques dans l'intérêt public.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

L'extraction finale et intégrale des données cliniques est réalisée par le data manager à partir du module d'extraction de CleanWEB
Les données seront archivées chez le promoteur selon la règlementation en vigueur pendant 15 ans après la fin de la recherche. Une copie de la base est sauvegardée sur le serveur.
Les métadonnées seront archivées sur EGA si la patiente ne s'y est pas opposée après information éclairée.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Les données cliniques et biologiques seront transmises par l'Unité de Recherche Clinique aux partenaires dans le format csv. Ce format csv peut être utilisé par tous les outils d'analyse de données.
Les données génétiques seront accessibles et exploitables (manipulation et visualisation) via des outils disponibles dans R, le logiciel libre. Les scripts et référentiels nécessaires à l'exploitation des métadonnées seront écrits, conservés dans le CLE et déposés sur GitHub pour archivage par les bioinformaticiens du laboratoire de recherche.
Des données seront stockées dans le service cloud GAIA de BForCure, localisé en Union Européenne, avec un accès tracé et sécurisé.

- L'accès se fait avec un JSON Web Token temporaire, un certificat SSL A+ (SSL Labs), ainsi que des accès bien identifiés par machines et plages de dates renseignées. Un premier accès à une machine est validé via un mécanisme d'authentification à facteurs multiples.
- Aucun accès direct à la machine n'est effectué depuis ce service cloud, la remontée des données est purement passive. Il n'y a donc aucun pare-feu à ouvrir, seul un accès internet au niveau de la machine est requis.
- Ce service cloud est déjà utilisé pour les études cliniques internes de BForCure, notamment pour avoir marqué CE leur kit SARS-CoV2.

Les outils statistiques seront développés sous le langage python, accessibles via des packages python, images docker ou services web, et le code source sera stocké sur gitlab.com, sur des dépôts privés.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Pour les métadonnées, EGA fournira un identifiant unique et stable.

---

6. Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Le responsable de la gestion des données est le Chef de projet RHU PrediMAP- APHP.
Les personnes impliquées dans le traitement des données cliniques:

- Technicien de recherche clinique pour la saisie des données clinique
- Attaché de Recherche Clinique : gestion et traitement des queries
- Datamanagement programmation de la base de données / eCRF, pour l'extraction, fusion des bases et le nettoyage de la base finale (contrôles de cohérence) vérification de la qualité des données
- Statisticien pour l'analyse et l'archivage de la base.
- Responsable de l'unité de recherche clinique
- Chef de projet promotion APHP pour les demandes réglementaires (CNIL…) en lien avec la recherche clinique
- Délégué à la protection des données pour rédaction et supervision des contrats à mettre en place entre les partenaires pour le transfert des données.

Données biologiques

- Chef de projet pour les données du laboratoire : analyse, archivage et partage des données
- Chercheurs, ingénieurs, doctorants pour la production, analyse
- Bioinformaticiens : gestion et traitement des données, vérification de la qualité des données, production des métadonnées,

Métadonnées : DAC (Data Access Consortium) : Chef de projet de laboratoire.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

- 390 000 euros pour chef de projet RHU 60 mois ETP.
- 120 000 euros pour la supervision des aspects techniques de la gestion des données de la recherche clinique.
- 280 000 euros pour les chefs de projet en charge de la supervision des aspects réglementaires de la recherche (chef de projet promotion, délégués à la protection des données
- 1 034 000 euros pour les techniciens de recherche clinique chargés de la saisie des données cliniques : 235 mois ETP
- 250 800 euros pour l'attaché de recherche clinique (gestion et traitement des queries) : 57 mois ETP
- 71 000 euros pour le datamanagement : programmation de la base de données / eCRF, pour l'extraction, fusion des bases et le nettoyage de la base finale (contrôles de cohérence) vérification de la qualité des données; 15 mois ETP
- 161 076 euros pour l'analyse et l'archivage de la base : PhD étudiant en épidémiologie (EPOPé), Ingénieur Statisticien (expérience 3-5 ans) et 2 stagiaires, M2 (EPOPé), 60 mois ETP.

  Soit un budget total de 2 306 876 euros pour l'ensemble de la gestion des données.