

Collecteur Analyseur de Données Projets de recherche

Mise en œuvre

Fonctionnement des bulles sécurisées

31 mars 2025 - bioinfodiag



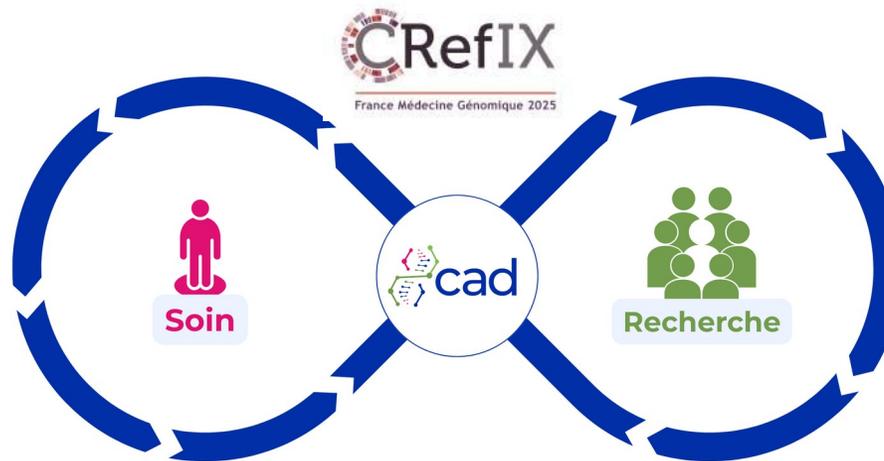
Infrastructures du plan

SeqOIA, Auragen, CRefIX, CAD

Le Plan France Médecine Génomique (PFMG2025) a pour objectif d'implémenter la Médecine Génomique en France dans un continuum soin-recherche. Il s'appuie sur 3 piliers principaux :

Un Centre de Référence, d'Innovation, d'expertise et de transfert (CRefIX)

Un réseau de laboratoires de séquençage génomique (LBM-FMG pour *Laboratoire de Biologie Médicale France Médecine Génomique*)



Un centre national de Collecte et d'Analyse des Données (CAD)



Collecteur Analyseur de Données

Quelques dates



- Appel à soumission de projets en 2022
- Recrutements des équipes courant 2024
- Autorisation EDS - juin 2024
- Données des LBM / infrastructure / premiers projets ouverts en cours

QU'EST-CE QUE LE CAD ?

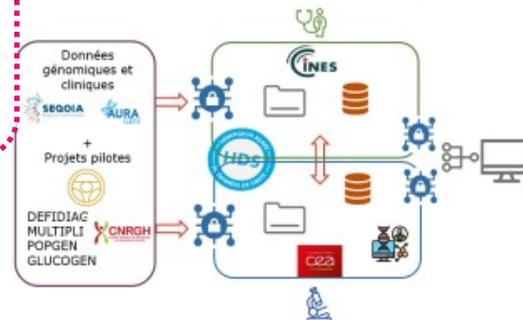
Le CAD est une **infrastructure nationale** portée par la puissance publique, notamment pour garantir la **souveraineté des données**, l'intérêt général et l'**encadrement éthique** de l'utilisation des données. Il offrira à la communauté scientifique un **accès sécurisé au catalogue de données génomiques issues du PFMG**, un éventail d'applications logicielles métiers, des **facilités de stockage et de calcul intensif**

Le CAD sera hébergé au **Très Grand Centre de Calcul (TGCC) du CEA** en Ile-de-France et au **Centre Informatique National de L'Enseignement Supérieur (CINES)** à Montpellier. Le CAD sera alimenté en continu par les **LBM-FMG (AURAGEN et SeqOIA à ce jour)** et recevra également les données provenant des **4 projets pilotes du PFMG2025 (MULTIPLI, DEFIDIAG, GLUCOGEN et POPGEN)**

Le CAD s'appuie sur plusieurs groupes de travail et un **Comité Scientifique et Ethique (CSE)**. Ce comité pluridisciplinaire, a pour rôle de s'assurer que les projets de recherche ayant accès aux données respectent un certain nombre de **critères et éthiques définis**

CAD <> Soins

- Ré-analyse des données dans la poursuite de la démarche diagnostique



CAD <> Recherche

- Réutilisation des données pour la recherche

LES SERVICES DU CAD

- **Catalogue** : collecte/annotations des données génomiques du PFMG
- **Appariement** avec les données cliniques (si besoin)
- Accès aux données dans des « bulles informatiques » sécurisées au service des chercheurs
 - Mise en place d'espaces sécurisés par projet de recherche
 - transfert des données PFMG utile pour le projet de recherche
 - possibilité pour les chercheurs d'y apporter leurs propres données et outils d'analyse
- **Stockage et ressources en calcul** dimensionnées aux besoins du projet
- Portail et cellule d'**accompagnement** de projets / **formation**

Projets de recherche

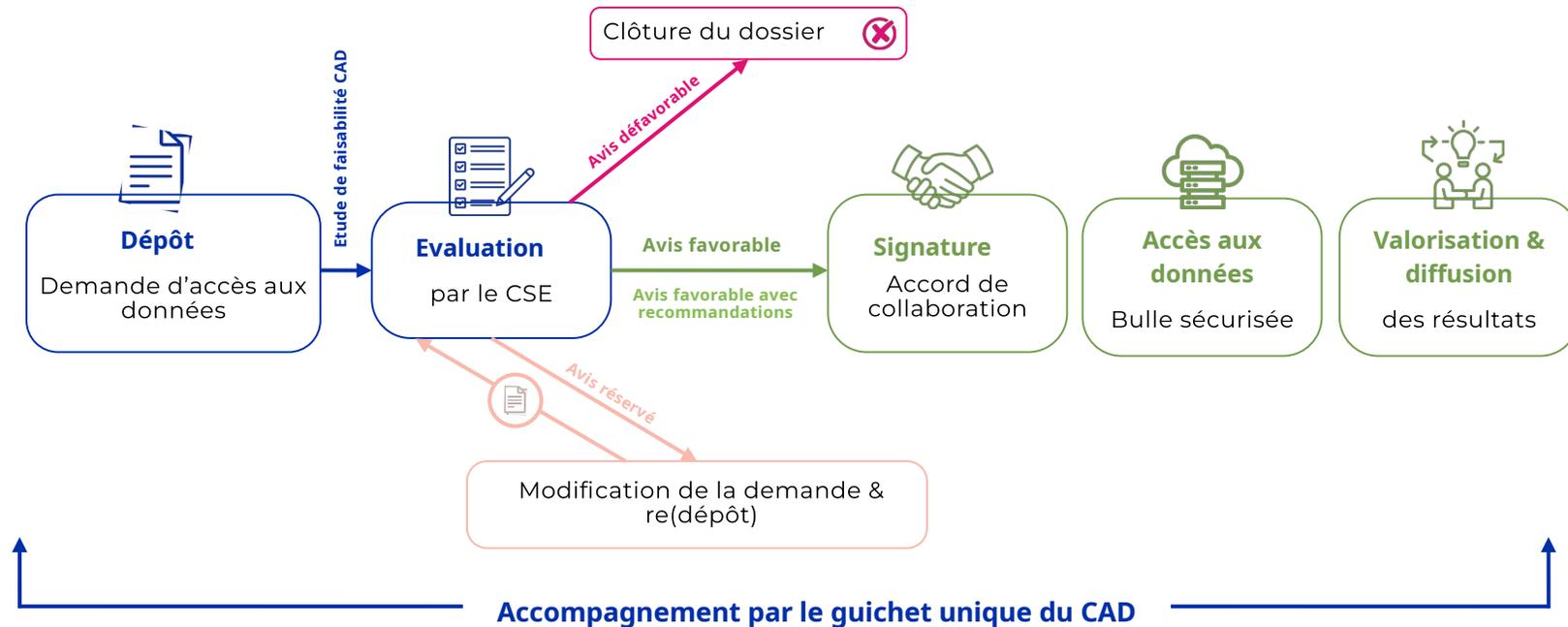
Cellule d'accompagnement des projets

Projets de recherche

- Conseil Scientifique et Ethique
- Données des Laboratoires de Biologie Médicale
- Bulle sécurisée - outils / données

Vers les "nouveaux diags"

Accompagnement de toutes les étapes



Comment déposer une étude?

<https://pfm2025.aviesan.fr/acces-aux-donnees/>

- Données générées par les LBM ?
- Demande d'accès : etude@genomecad.fr

Données d'activité du PFMG2025
 Par préindication au 31 décembre 2024

Dossier de demande d'accès au Collecteur Analyseur de Données (CAD) en vue de la réutilisation de données produites dans le cadre du PFMG2025 pour des projets de recherche

Version mai 2023 [Télécharger](#)



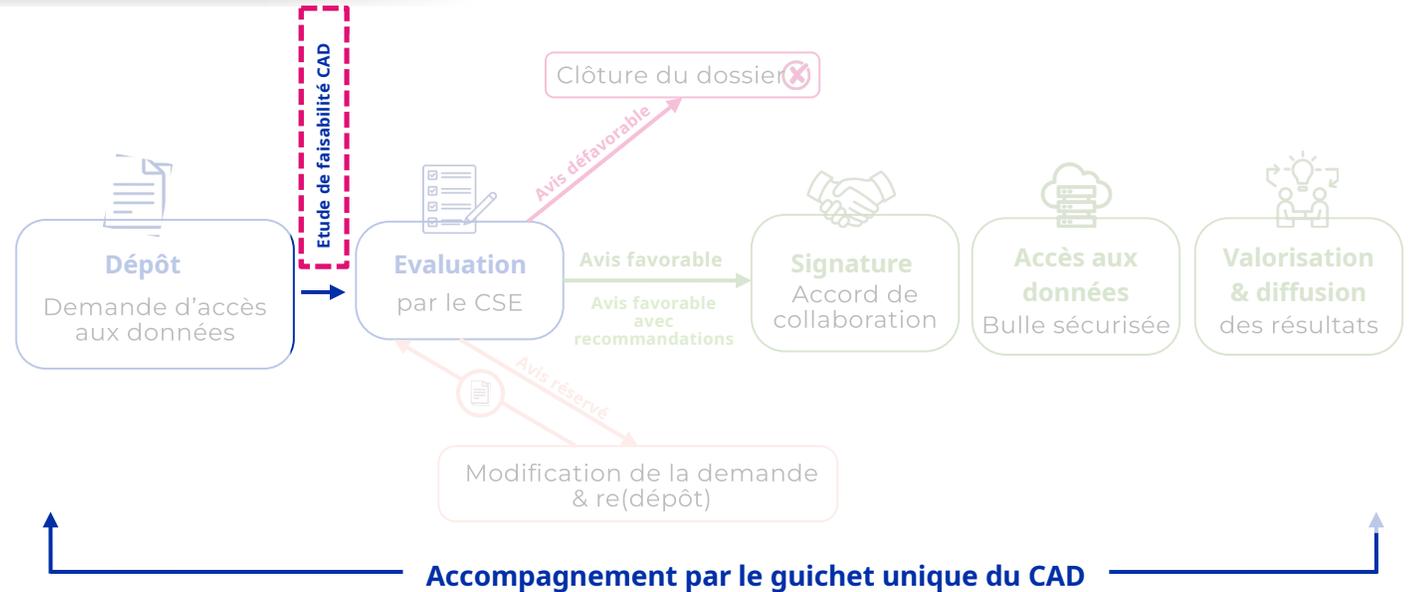
Données d'activité du PFMG2025 au 31 décembre 2024

	Nb de prescriptions validées en RCP-FMG	Nb de dossiers complets	Nb de dossiers clos
Angiodèmes bradykiniques héréditaires	3	3	2
Anomalies du développement, syndromes malformatifs et syndromes dysmorphiques sans déficience intellectuelle	9 098	8 123	6 040
Anomalies sévères de la différenciation sexuelle d'origine gonadique et hypothalamo-hypophysaire	186	141	96
Aplasies et hypoplasies médullaires	65	59	52

Mise en place "faisabilité CAD" : pourquoi?

Avant mise en place d'une "faisabilité CAD"

- Conformité réglementaire ?
- Est ce que les données demandées sont générées ? Disponibles au sein du CAD ?
- Ce qui est décrit dans le projet, est-ce faisable au CAD ? Outils ... infra ... ?



- Proposition d'accompagnement Faisabilité CAD
- Structurer pour limiter les projets qui stagnent
- Accompagner CSE

Faisabilité CAD : retours aux membres CSE & porteurs des projets

Collecteur Analyseur de Données

1. Conformité réglementaire
Évalués par l'équipe opérationnelle CAD

	OUI	NON
L'étude déposée contient des données propres identifiantes.	<input type="checkbox"/>	<input type="checkbox"/>
L'étude déposée contient des données propres indirectement identifiantes.	<input type="checkbox"/>	<input type="checkbox"/>
L'étude déposée contient des données propres anonymisées.	<input type="checkbox"/>	<input type="checkbox"/>
Conformité au référentiel "entrepôt de santé CAD" si autorisation supplémentaire de la CNIL.	<input type="checkbox"/>	<input type="checkbox"/>
Informations relatives à l'appariement des données.	<input type="checkbox"/>	<input type="checkbox"/>
L'équipe projet prévoit la rédaction d'une Analyse d'Impact à la protection des données (AIPD).	<input type="checkbox"/>	<input type="checkbox"/>
L'équipe projet prévoit une déclaration pour le traitement de l'analyse prévue.	<input type="checkbox"/>	<input type="checkbox"/>
L'équipe projet prévoit l'accès à la partie analyse à un responsable de traitement.	<input type="checkbox"/>	<input type="checkbox"/>

2. Méthodologie
Expression du besoin en données et fichiers

	PRESENTS	ABSENTS	Commentaire
Les dossiers d'intérêt pour réaliser l'étude répondent à ces critères : <i>Renseigner par exemple la préindication / les informations cliniques / les structures familiales demandées.</i>	<input type="checkbox"/>	<input type="checkbox"/>	
Au regard des informations actuellement disponibles au sein du CAD, nombre de dossiers	<input type="checkbox"/>	<input type="checkbox"/>	

3. Faisabilité
Évaluée par l'équipe opérationnelle CAD

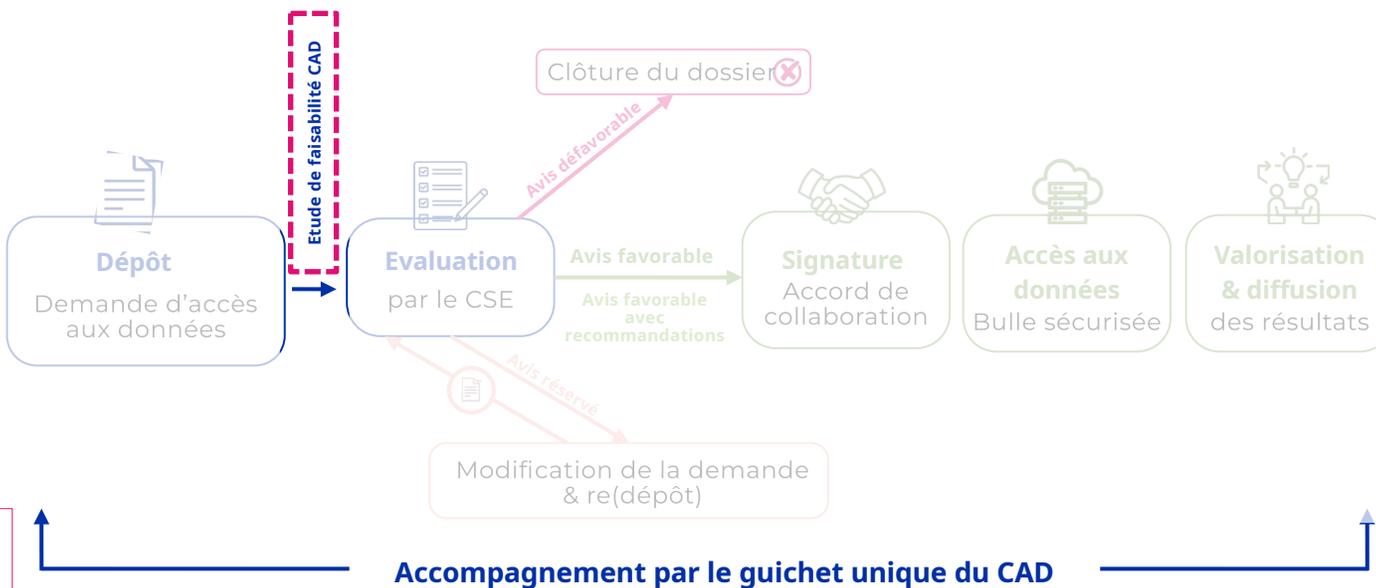
	OUI	NON	Commentaire
Les ressources informatiques demandées correspondent aux traitements prévus. <i>Si non applicable, préciser.</i>	<input type="checkbox"/>	<input type="checkbox"/>	
Des informations concernant les processus sont fournies de manière "standard". <i>Ex : 24 core cpus / 32 Go de RAM.</i>	<input type="checkbox"/>	<input type="checkbox"/>	
Description de l'enchaînement des analyses (gestionnaire de données) présente.	<input type="checkbox"/>	<input type="checkbox"/>	

CONCLUSION

Faisabilité confirmée Précisions à apporter

Conclusions de l'équipe opérationnelle CAD :

DPO de traitement : _____
Nom, prénom et email du DPO du responsable de traitement : _____



Evaluation par le Conseil Scientifique et Ethique - CSE

Fiche d'évaluation de projet

1. Objectifs, Méthodologie et données de l'étude

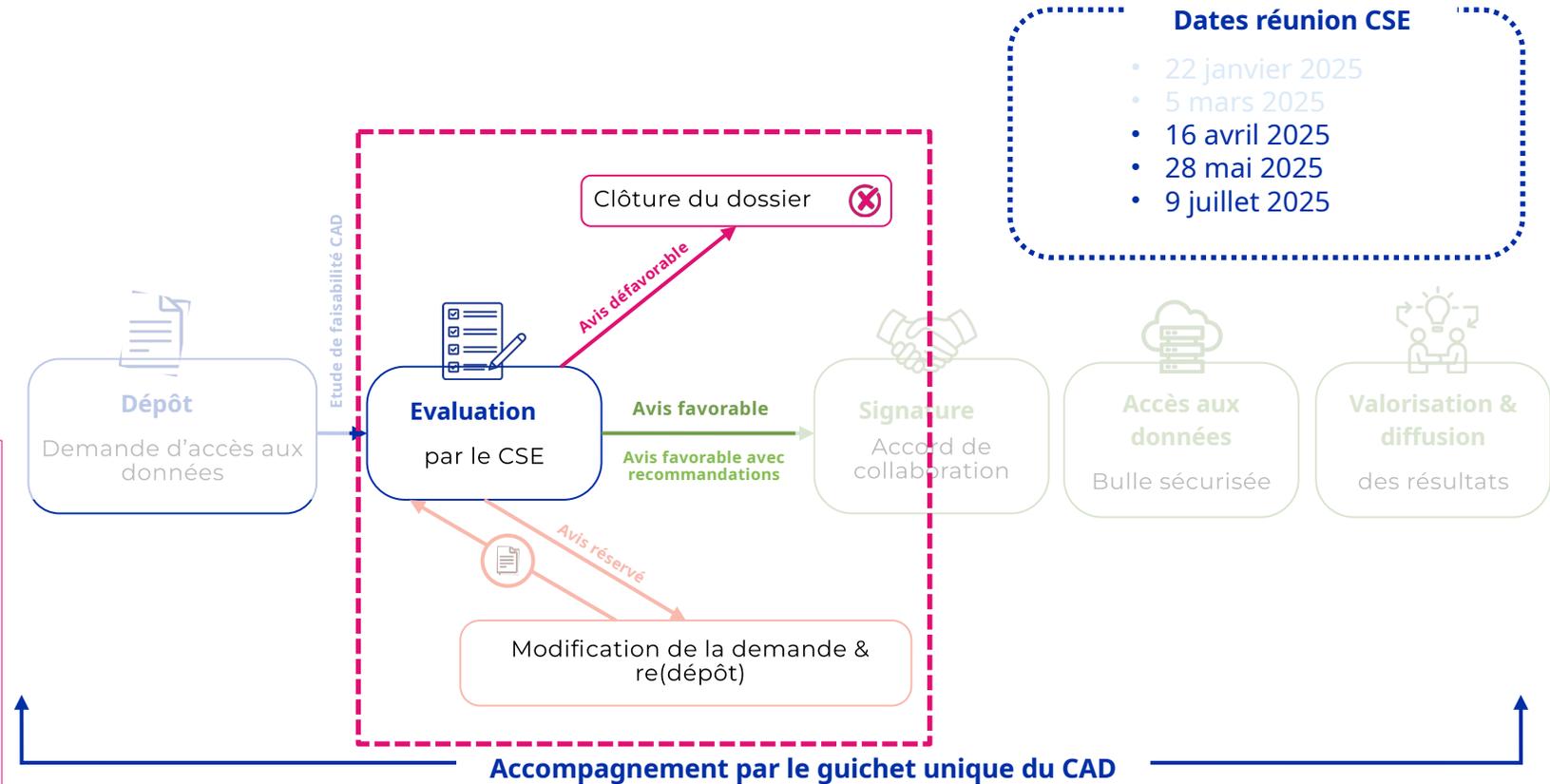
	OUI	NON	N/A
Le contexte scientifique, médical, épidémiologique est clairement exposé.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
La justification/intérêt de l'étude est démontré(e).	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Avis proposé par le rapporteur

Sur la base des questions précédentes :

	OUI	NON
L'étude est conforme à l'éthique*	<input type="checkbox"/>	<input type="checkbox"/>
L'étude présente un intérêt scientifique et/ou social	<input type="checkbox"/>	<input type="checkbox"/>

AVIS FAVORABLE
 AVIS FAVORABLE avec recommandations
 AVIS RÉSERVÉ (nouveau passage en CSE nécessaire)
 AVIS DÉFAVORABLE (nouveau dépôt de projet nécessaire)



2 rapporteurs par projets discuté, discussion avec tout le CSE

Les CSE de 2022 à 2024 : 17 projets favorables

Évaluation CSE	2022	2023	2024
Défavorable	-	1	-
Favorable avec recommandations			
Favorable	3	8	6
Réservé	5	13	8



PROJETS DE RECHERCHE

Accueil / Projets de recherche

Vous trouverez ici les résumés grands publics des projets de recherche liés au PFMG2025 :

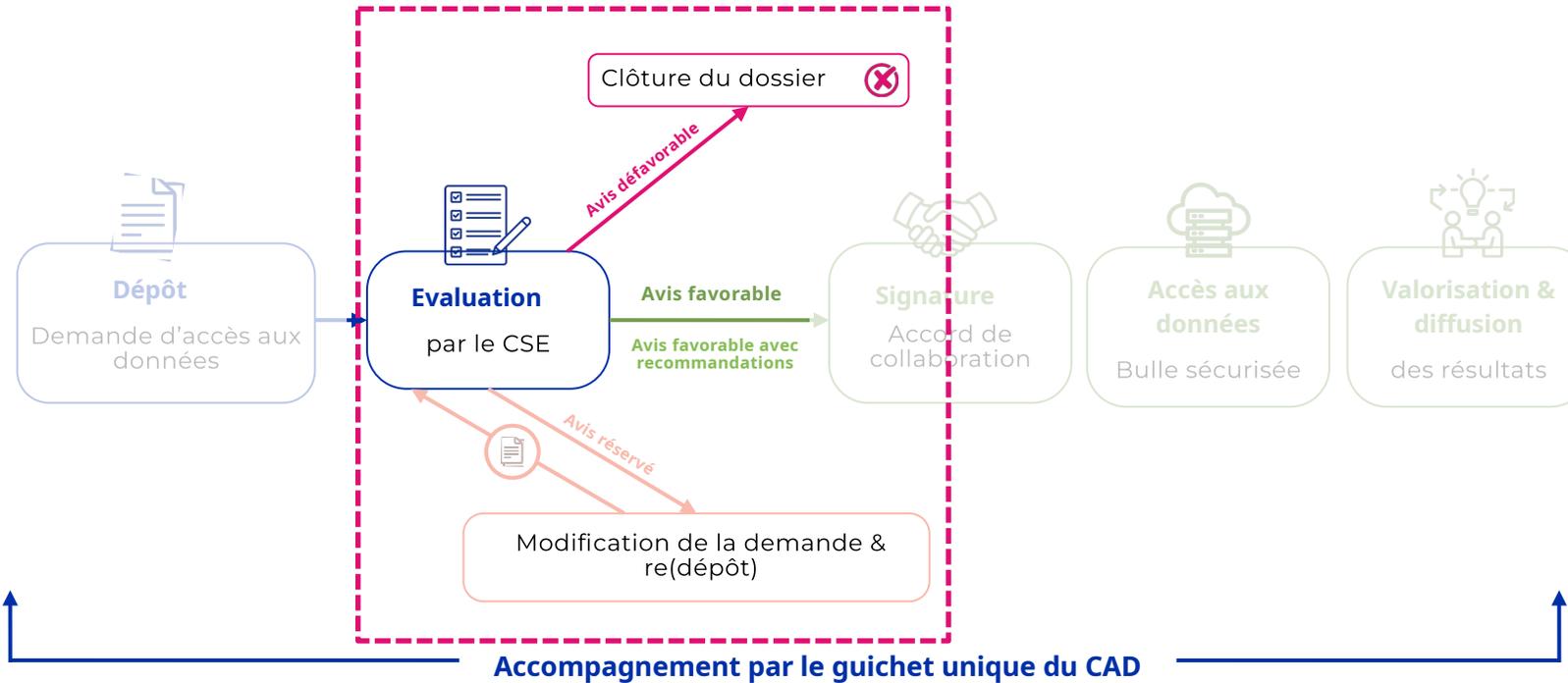
- Projets sur les aspects médico-économiques :
 - Projet de recherche [SONCO \(Séquençage Oncologie Cohorte\)](#)
- Projets ayant fait une demande au CAD, validés par le CSE et dont la réutilisation des données n'a pas encore commencé :
 - Projet de recherche [PAAE \(PFMG ré-Analyse des Atouts Génétiques pédiatriques séquentielles Génomiques\)](#)
 - Projet de recherche [REACT \(Recherche d'Atouts Diagnostiques dans les cardiomyopathies\)](#)
 - Projet de recherche [GEDI \(Génétique de la Déficience Intellectuelle et Imagerie\)](#)
 - Projet de recherche [ACE-ESMART \(Ancestry-informed ESMART: Ancestry-informed Molecular Anomalies in Relapsed or refractory Tumors\)](#)
 - Projet de recherche [PANDORA \(PANDORA: Personalized and Immunotherapeutic of sarcomas\)](#)
 - Projet de recherche [DenovoRank](#)
 - Projet de recherche [Identification et validation de nouveaux gènes et mutations dans les maladies neuromusculaires](#)
 - Projet de recherche [PRA \(PRA: Recherche de causes des sarcomes moléculaires chez les patients atteints de cancers rares : cohorte prospective nationale\)](#)
 - Projet de recherche [EAG-ADDIMN \(EAG-ADDIMN \(EAG-Analyse des Génomes pour les Anomalies du Développement, les Déficiences Intellectuelles de causes rares, les Maladies Mitochondriales et Neurogénétiques\)\)](#)
 - Projet de recherche [MESOSCREEN \(Genomic and transcriptomic analyses identify a prognostic gene signature and predict response to therapy in epithelial mesothelioma\)](#)
 - Projet de recherche [PAS \(PFMG ré-Analyse Surdités\) : Identification de nouveaux déterminants génétiques dans les surdités précoces](#)
 - Projet de recherche [MICROC : Exploiter les données génomiques du CAD pour améliorer le diagnostic des patients atteints de microcéphalie](#)
 - Projet de recherche [Gènes et régions non-codantes impliqués dans les anomalies développementales structurales du pont et du cervelet](#)
 - Projet de recherche [Identification et validation de nouveaux gènes et mutations dans les maladies neuromusculaires](#)
 - Projet de recherche [Développement d'approches informatiques basées sur le Deep Learning pour l'évaluation conjointe des CNV \(Copy Number Variant\) et des SNPs \(Single Nucleotide Polymorphism\) identifiés par séquençage de génome dans une cohorte de patients présentant une déficience intellectuelle sous Thyroïdisme](#)
 - Projet de recherche [PA-DRSP \(PFMG ré-Analyse DRSP\) Identification des causes génétiques responsables de maladies rétinienne de l'enfant](#)
 - Projet de recherche [Maladies osseuses constitutionnelles : éluder l'errance diagnostique et comprendre la variabilité phénotypique](#)

Avenir 2025
Portail de transparence
projets CSE

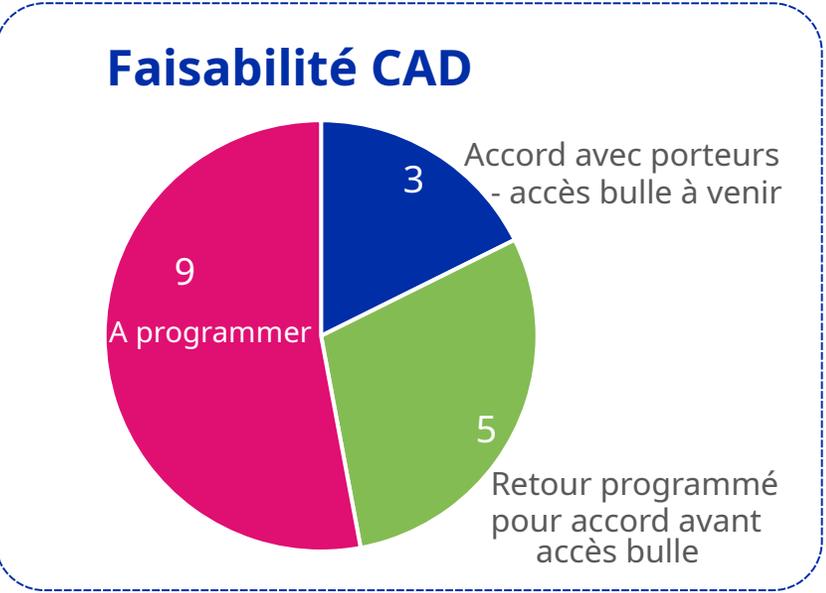
- Faisabilité CAD = atouts CSE
- Action pour mise en oeuvre des bulles

<https://pfm2025.aviesan.fr/projets-de-recherche-2/>

Faisabilité CAD en pratique



- Faisabilité pour acter les outils / données à mettre en accès dans les bulles sécurisées
- **2025 : Mise en place des accès aux bulles**



Transfert des données des LBM



Soin & transfert des données

- de prescription
- de séquençage (dossiers interprétés)



Projet X

- Données laboratoires pseudonymisées
 - o Critères cliniques
- Données propres
- Outils du catalogue / propres au projet
- Puissance / stockage demandé
- Jeu de référence asymptomatique



Projet Y

- Données laboratoires pseudonymisées
 - o Critères cliniques
- Données propres
- Outils du catalogue / propres au projet
- Puissance / stockage demandé
- Jeu de référence asymptomatique

Validations internes avant mise en œuvre projets

Réglementaire

Conforme "Entrepot Santé CAD"
Contractualisation
Analyse d'impact relative à la protection des données
Déclaration [Health Data Hub](#)
Habilitation membres du projet



**Formalisation / procédures en
cours de rédaction**

Données

Faisabilité CAD - précision du besoin des porteurs de projet
Génération des données sur les plateformes
Transfert des données PFMG dans la bulle
Données propres attendues dans la bulle
Vérification des données transférées dans la bulle

Logiciels / bulle

Faisabilité CAD : précision du besoin logiciels
Fiche technique bulle sécurisée
Configuration espace logiciels
Ressources de calcul et stockage
Données de référence dans la bulle
Bulle sécurisée HDS accessible

Ressources

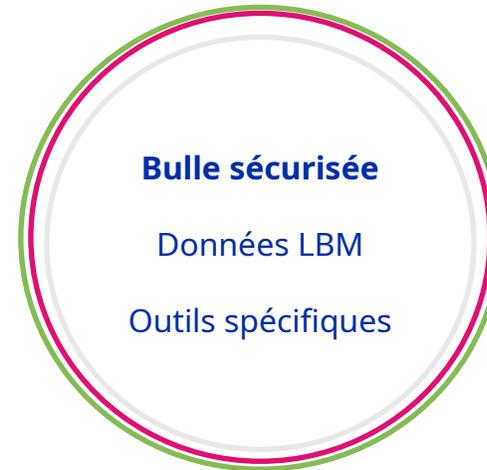
Bioinformaticien
Espace collaboratif CAD

Trois grandes étapes de mise en œuvre Projets avec données des LBM



Projets "structurants"

Disponibilité des équipes pour rendre la bulle fonctionnelle et permettre au projet de se réaliser



Population de référence Défidiag

- SNV
- CNV
- SV



Besoin d'ajouter des outils ? Des données ?

Ex :

- Visualisation des données
- Dictionnaire HPO



Catalogue CAD

- Outils / annotations
- Variations



Mise en place d'un "**welcome pack**" défini en fonction des besoins pointés par projets structurants

☑ Sera incrémenté avec le temps

Catalogues qui vont s'incrémenter



Dictionnaire des termes HPO
Outil de visualisation IGV
Python

Annotations structurales et fonctionnelles

- Annotations structurales - gencode / RefSeq
- Annotations géniques - HGNC / OMIM
- Données génomiques - TCGA
- Annotations cliniques et phénotypiques - Orphanet / HPO
- Annotations populationnelles - 1000 genomes / gnomAD exome / genome
- Annotations pathogénicité - COSMIC / ClinVar
- Annotations d'impact - CADD / dbNSFP
- ... en cours!



Outils d'appel des variations

- samtools 1.19
- snpeff 5.2
- snpsift 5.2
- finsurf a2c989c
- bedtools 2.31.1
- bedops 2.4.41
- R 4.3.2
- epacts 3.4.2
- spip 2.1
- vep 111
- ... en cours

Portail CAD : objectif transparence et requêtage des données présentes dans l'EDS / des catalogues

	Nb de prescriptions validées en RCP-FMG	Nb de dossiers complets	Nb de dossiers clos
Angiodèmes bradykiniques héréditaires	Aujourd'hui : Données fournies par le PFMG + faisabilité CAD	3	2
Anomalies du développement, syndromes malformatifs et syndromes dysmorphiques		8 123	6 040
Anomalies sévères de la différenciation sexuelle d'origine gonadique et hypothalamo-hypophysaire		141	96
Aplasies et hypoplasies médullaires		65	59

A construire :

Informations requêtable : dossiers accessibles pour projet

- Préindication?
- Informations cliniques ?
- Structures familiales
- ... à définir en fonction des besoins projets

- Catalogue "variations" ?
SNV CNV SV ...

Quels outils sont installés ?

- Variations du nombre de copies ?
Détail
- Fusions ?
- ..

Quelles annotations peuvent être utilisées ?

- Annotations structurales ?
Détail
- Annotations fonctionnelles ?

Projet favorable CSE – dans la bulle ?



Outils / ressources mises à disposition **"Welcome pack"**



Données populationnelles
disponibles au CAD (DEFIDIAG, ...)



Outils du catalogue qui répondent aux questions du projet

Outils spécifiques validés "faisabilité CAD".....

Ressources d'annotation du catalogue

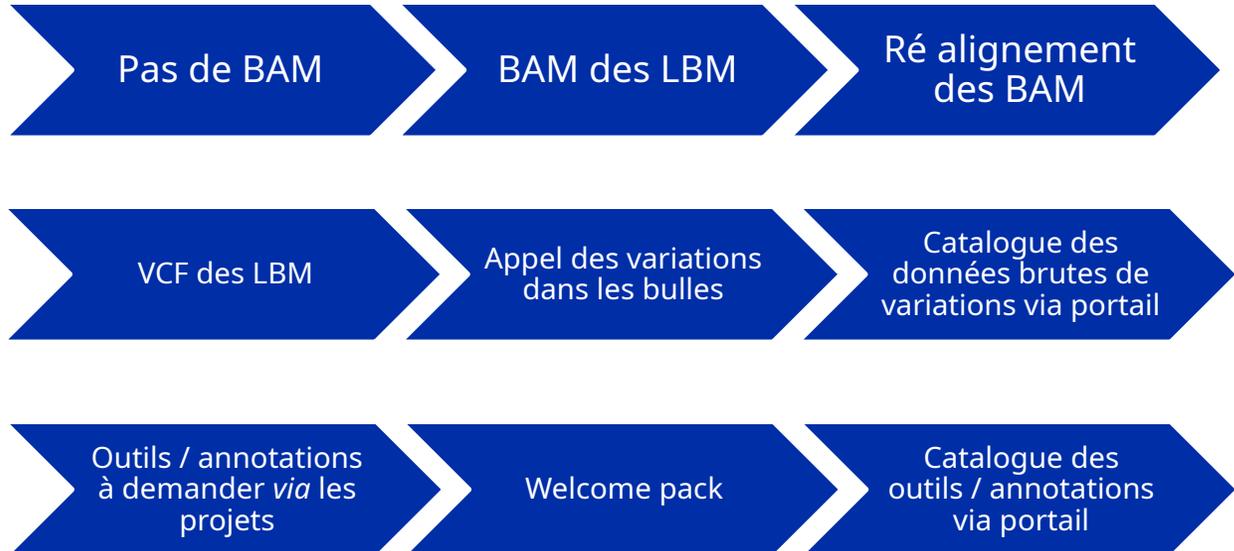
Annotations spécifiques validés "faisabilité CAD".....



Données des LBM – informations cliniques

Données propres des projets - déclaré dans le projet

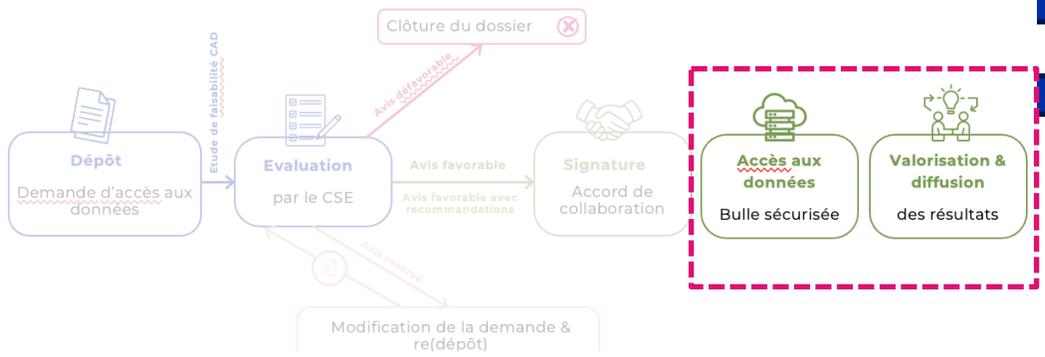
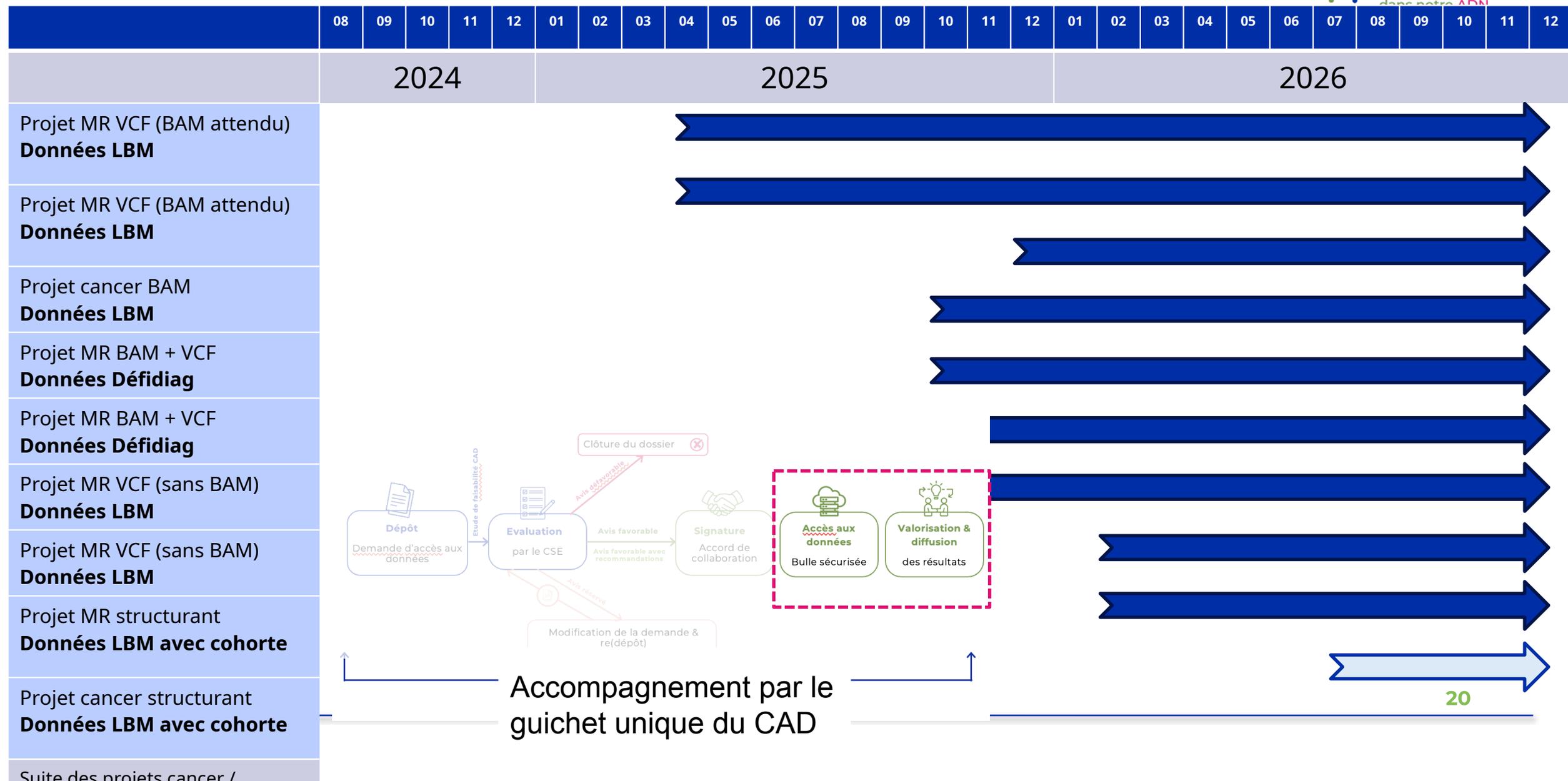
Fonctionnement actuel, et objectifs



Autonomie des porteurs à requêter les catalogues outils / annotations / variations

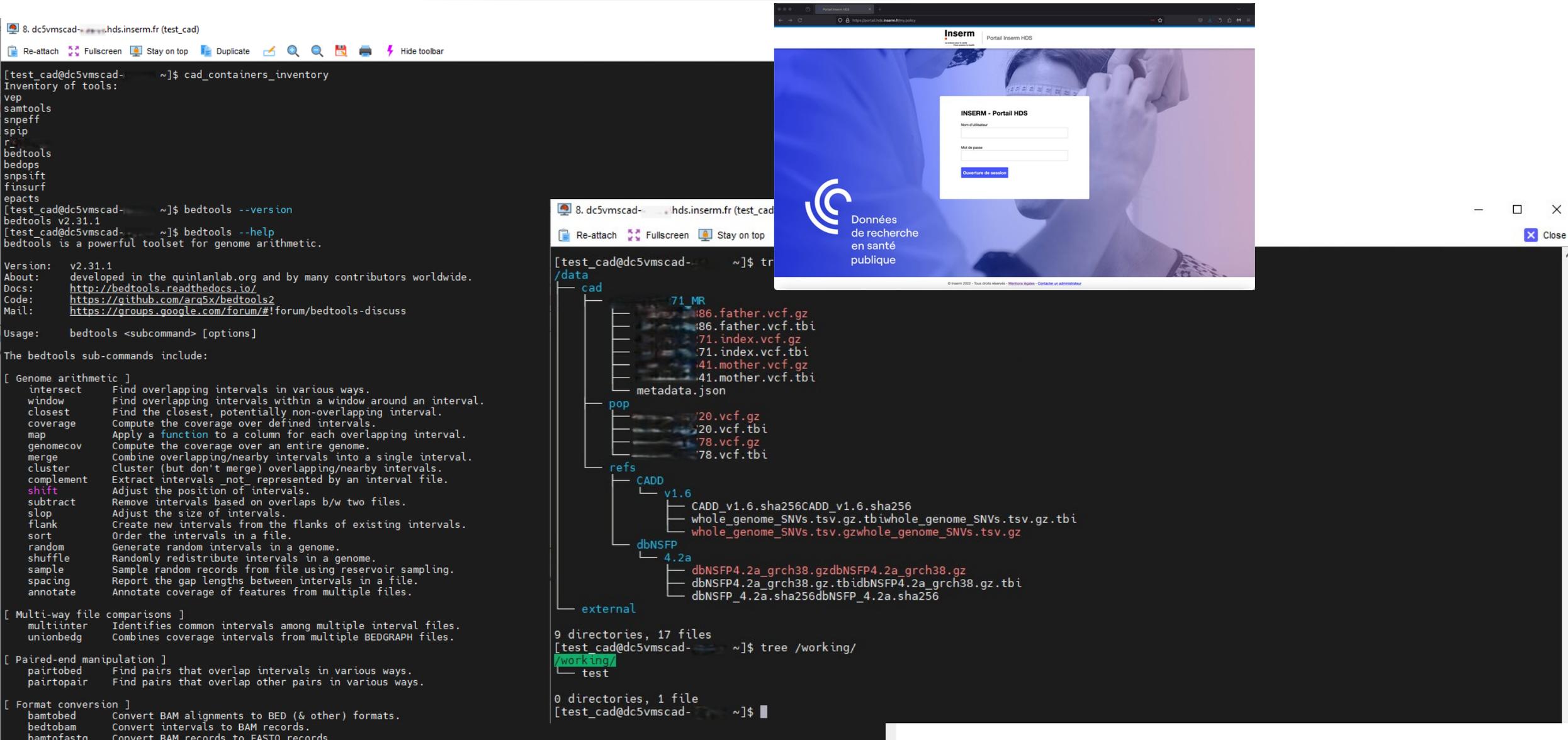
Processus automatisés

Accès aux données & valorisation



Accompagnement par le guichet unique du CAD

Bulle sécurisée



The image shows a terminal window on the left and a browser window on the right. The terminal window displays the following commands and output:

```
8. dc5vmscad-@hds.inserm.fr (test_cad)
Re-attach Fullscreen Stay on top Duplicate
[test_cad@dc5vmscad- ~]$ cad_containers_inventory
Inventory of tools:
vep
samtools
snpeff
spip
r
bedtools
bedops
snpsift
finsurf
epacts
[test_cad@dc5vmscad- ~]$ bedtools --version
bedtools v2.31.1
[test_cad@dc5vmscad- ~]$ bedtools --help
bedtools is a powerful toolset for genome arithmetic.

Version: v2.31.1
About: developed in the quinlanlab.org and by many contributors worldwide.
Docs: http://bedtools.readthedocs.io/
Code: https://github.com/arq5x/bedtools2
Mail: https://groups.google.com/forum/#!forum/bedtools-discuss

Usage: bedtools <subcommand> [options]

The bedtools sub-commands include:

[ Genome arithmetic ]
intersect Find overlapping intervals in various ways.
window Find overlapping intervals within a window around an interval.
closest Find the closest, potentially non-overlapping interval.
coverage Compute the coverage over defined intervals.
map Apply a function to a column for each overlapping interval.
genomecov Compute the coverage over an entire genome.
merge Combine overlapping/nearby intervals into a single interval.
cluster Cluster (but don't merge) overlapping/nearby intervals.
complement Extract intervals not represented by an interval file.
shift Adjust the position of intervals.
subtract Remove intervals based on overlaps b/w two files.
slop Adjust the size of intervals.
flank Create new intervals from the flanks of existing intervals.
sort Order the intervals in a file.
random Generate random intervals in a genome.
shuffle Randomly redistribute intervals in a genome.
sample Sample random records from file using reservoir sampling.
spacing Report the gap lengths between intervals in a file.
annotate Annotate coverage of features from multiple files.

[ Multi-way file comparisons ]
multinter Identifies common intervals among multiple interval files.
unionbed Combines coverage intervals from multiple BEDGRAPH files.

[ Paired-end manipulation ]
pairtobed Find pairs that overlap intervals in various ways.
pairtopair Find pairs that overlap other pairs in various ways.

[ Format conversion ]
bamtobed Convert BAM alignments to BED (& other) formats.
bedtobam Convert intervals to BAM records.
bamtofastq Convert BAM records to FASTQ records.
```

The browser window shows the Inserm HDS portal login page. The page title is "Inserm Portail Inserm HDS". The main content area features a large image of a person's hands holding a measuring tape. A login form is centered on the page with the following fields and buttons:

- Form title: INSERM - Portail HDS
- Field: Nom d'utilisateur
- Field: Mot de passe
- Button: Couverture de session

Below the login form, the text "Données de recherche en santé publique" is visible. The browser address bar shows "https://portal.hds.inserm.fr/portal".

Participation active de la bioinfo, bel atout!

Structuration de la mise en oeuvre des projets en 2025

- Infrastructure
- Premières données des dossiers LBM
- Mise en place de la qualité
- Mise à disposition de bulles sécurisées en cours
- Catalogue des outils / des données
- "Welcome pack" à définir (et incrémenter)
- Calendrier des livraisons

Perspectives: les projets "au fil de l'eau" avec catalogue requêtable

Merci !



Cellule accompagnement des projets

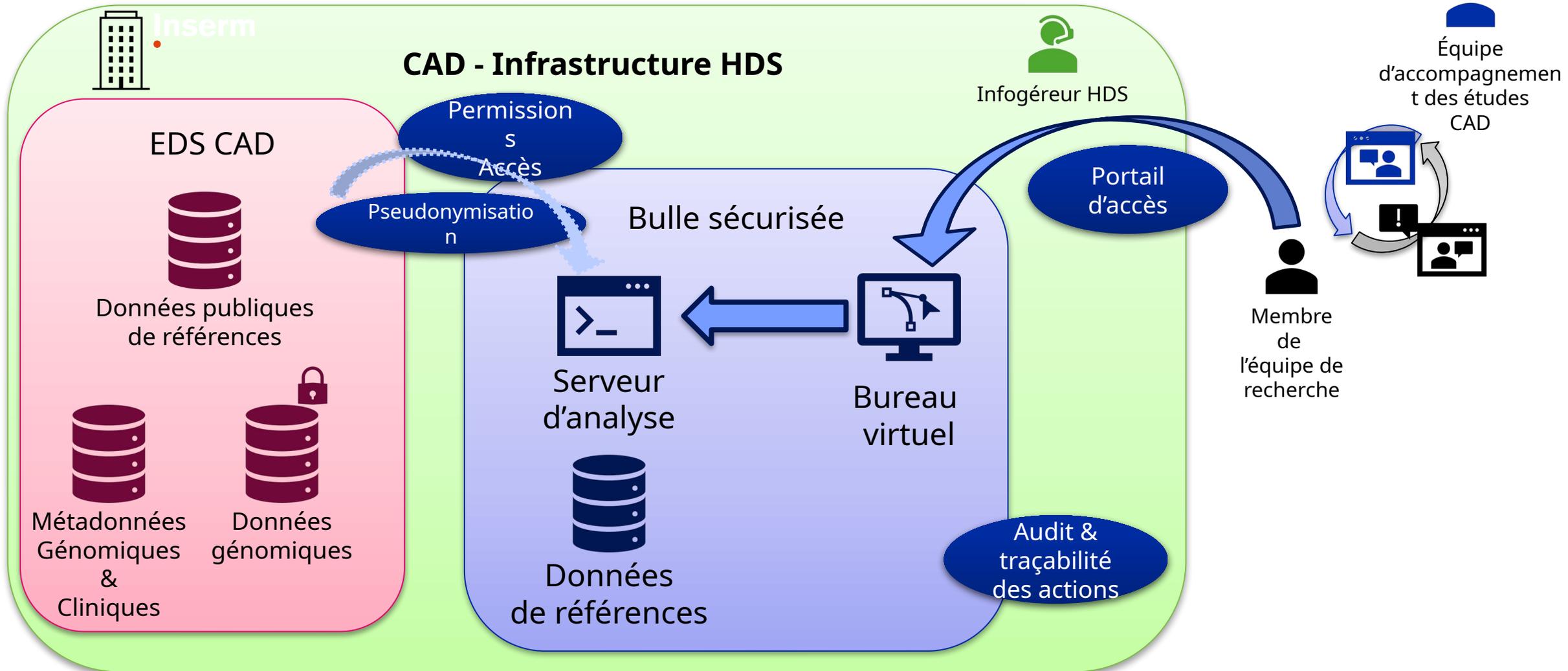
Adrien Josso Rigonato
Flavian Rique
Syrine Bouallegue
Cécile Meslier
Virginie Bernard



Contact : Virginie Bernard



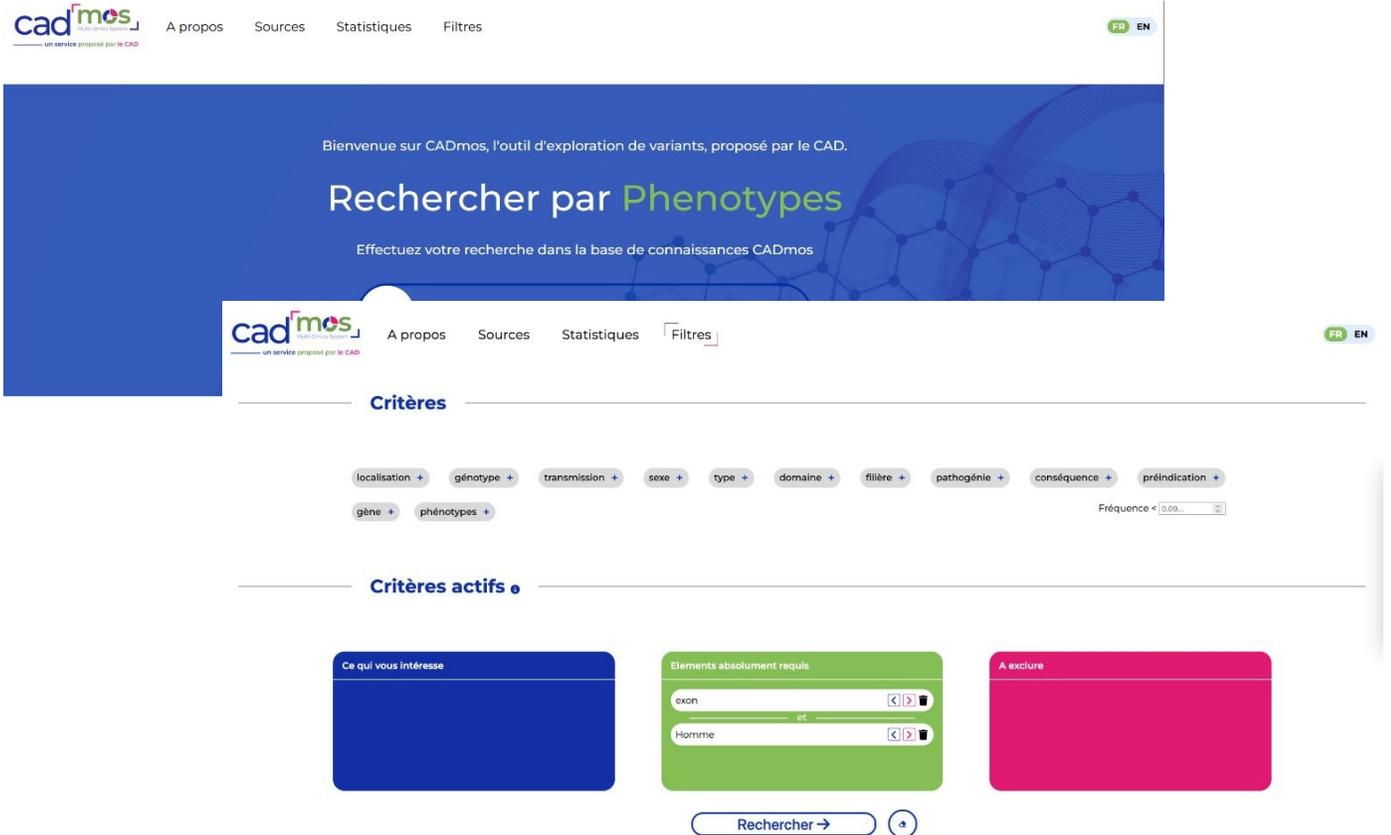
Environnement pour la recherche



Moins de projets en 2024 ? Anticipation de l'intérêt CADMOS

Évaluation CSE	2022	2023	2024
Défavorable	-	1	-
Favorable avec recommandations			
Favorable	3	8	6
Réservé	5	13	8

- Projets avec focus sur gène d'intérêt
- Variations d'intérêt avec clinique spécifique? >> CADMOS
- Ex. RNU4-2 & biais



The screenshot shows the CADmos website interface. At the top, there is a navigation bar with links for 'A propos', 'Sources', 'Statistiques', and 'Filtres'. The main header area features a blue background with the text 'Bienvenue sur CADmos, l'outil d'exploration de variants, proposé par le CAD.' and 'Rechercher par Phenotypes'. Below this, there is a search bar and a section for 'Critères' (Criteria) with various filters like 'localisation', 'génotype', 'transmission', etc. The 'Critères actifs' (Active criteria) section shows three panels: 'Ce qui vous intéresse' (empty), 'Elements absolument requis' (with 'exon' and 'Homme' selected), and 'A exclure' (empty). A 'Rechercher' button is at the bottom.