ArcaScience

BR-PREDICT

La première plateforme d'évaluation prédictive du bénéfice-risque

28/01/2026

ArcaScience en résumé

Date de création : 2018

Montants levés au total :

€**9M**

Dernier fundraising: **€4,5M**

en sept. 2025

Reconnue meilleure healthtech NLP Startup 2023[1]

10+ clients blue chip client utilisant notre solution

Révolutionner l'évaluation du bénéfice-risque pour la recherche clinique et les autorités de santés, **grâce à des données BRA exhaustives (100b) & 24 IA fiables,** pour optimiser les essais en temps réel.

[1] : https://www.startus-insights.com/innovators-guide/natural-language-processing-startups/

Une équipe expérimentée et pluridisciplinaire

Fondateurs

Romain CLEMENT
CEO & Co-Fondateur

Expert en données cliniques et en stratégie d'innovation

Patrick Pierre
Co-Fondateur et superviseur technique infrastructure

Ingénieur expérimenté en infrastructures numériques

Laurent Romary
Co-Fondateur, scientific advisor

Directeur de recherche INRIA,
Co-inventeur LLM moderne

Equipe dirigeante

Vassili Soumelis
Chief Medical O.

Immuno-oncologue, hématologue, ancien CMO Owkin, expert AI

Julien Dufour
Chief Business O.

Spécialiste du développement commercial et des stratégies marché dans la santé

Jean-François Arbona
CTO

Expert en architecture cloud et en développement de produits digitaux

Equipe opérationnelle

Equipe active

+1 administratif
+1 prestataire de recrutement
+3 stagiaires

Recrutements prévus pendant le projet

	2026	2027	2028	2029	Total
R&D	6	4	13	4	27
Industriels	2	0	0	4	6
Commercial	2	1	7	4	14
Total	10	5	20	12	47

Un conseil scientifique, d'envergure mondiale

Expertise Médicale

Pr. Clemens M. Schirmer, MD, PhD,
Geisinger Hospital

Stéphane Rouault, PhD
ArgenX

Pr. Alexis Brice, MD, PU-PH
Institut du Cerveau

Pr. Jean Lopategui, MD, PhD
Cedars Sinai Hospital

Expertise Technique

Laurent Romary, PhD
INRIA

Geneviève Laurans
Quotium

Serge Bauin, PhD
CNRS

Emmanuel Capitaine, MD
Vivalto Santé

Expertise pharmaceutique

Kerry Coffee
Weil Cornell Medicine

Philippe Peyre, PhD
Sanofi

Hal Lavender
CapGemini

Béatrice du Sordet
AplusA

Fabien Lanteri
Genopole

Développement clinique d'un médicament : des investissements massifs pour un taux de réussite trop faible

Les essais cliniques depuis le début du XXIème siècle :

Un investissement moyen estimé entre 1 et 2 Milliards de dollars pour chaque nouvelle molécule commercialisée pour 92% d'échecs.

Inclusion de milliers de patients malades et de sujets sains

Causes des échecs

40 à 54 % des échecs sont attribués à un manque d'efficacité

17 à 30 % sont dus à un problème de toxicité

Et si nous pouvions mieux anticiper les résultats, en amont du développement clinique?

Pratiques actuelles et limites de l'évaluation du bénéfice-risque (BR)

L'évaluation du BR détermine si les bénéfices attendus d'un médicament justifient les risques liés à son usage.
Elle est une étape décisive du processus réglementaire, conduite par les industriels pharmaceutiques et les autorités réglementaires à des moments clés du cycle de vie d'un médicament.
En clinique, elle coûte **~$13,2b/an. Ce montant ne prend pas en compte les 280k nouvelles molécules candidates chaque année**.

Ce processus est réalisé par les équipes internes des laboratoires pharmaceutiques ou délégué à des CRO spécialisées.

- Une approche **structurée** d'évaluation des bénéfices et des risques n'est pas encore largement adoptée
- Evaluations très coûteuse en temps (6/18 mois) et en temps homme (4 à 5 consultants).

- Processus chronophage
- Hétérogénéité des pratiques entre phases précoces et tardives
- 90 à 99% des données quanti/quali de BR non-inclues
- Haute hétérogénéité des données
- Capacité prédictive réduite
- Processus coûteux : de 500k à 1,2M € en fonction de la phase

Aujourd'hui, ce type d'évaluation reste insuffisamment prédictive et stratégique en phase précoce. Une approche structurée et anticipée permettrait d'optimiser les décisions go/no-go et de réduire les risques d'échecs coûteux en phase tardive

Des besoins forts dans l'optimisation et l'automatisation de l'évaluation du bénéfice-risque précoce

Renforcer la précision des analyses

Harmoniser l'exploitation des données hétérogènes massives

Améliorer l'impact social

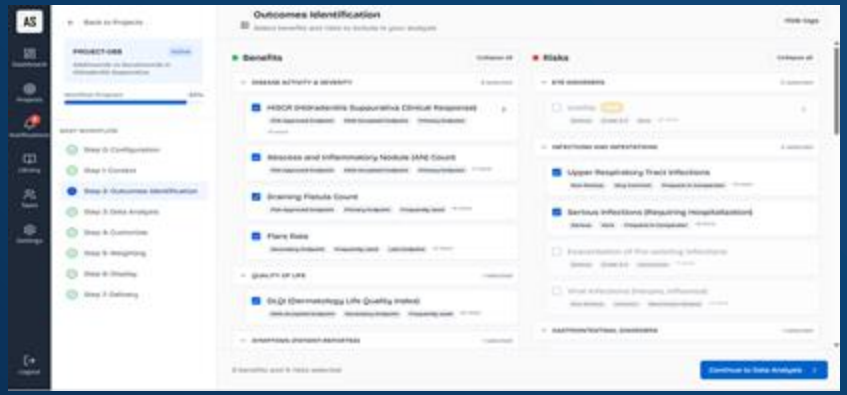
Réduire la durée et les coûts

Agir en amont

ArcaScience : la première plateforme permettant de construire l'analyse exhaustive et automatique du bénéfice-risque clinique

Bientôt accessible pour le screening de candidats précoces

Notre plateforme :
1- Extrait et collecte massivement les données cliniques et biomédicales issues de sources fermées et ouvertes depuis les serveurs du client
2- Analyse automatiquement le profil bénéfice-risque d'un candidat médicament.
3- Compare le candidat au traitement de référence pour identifier ses forces et ses faiblesses
4- Préconise un positionnement optimal par rapport aux options thérapeutiques existantes.



Ses avantages uniques :

Parcours utilisateur structuré aligné sur le cadre BRAT

Capacité d'intégration dynamique de toute source sémantique pertinente

Automatisation d'extraction, standardisation, normalisation & classification

Workflow technique populé grâce à 24 models d'IA auditables, fiables, traçables

Bénéfices promis par la solution ARCASCIENCE : **Construire le profil bénéfice-risque**

Un gain de temps et de coûts massif

125 k€ **VS** ≈ 500k/1,2 M€

Few minute **VS** 18 mois.

Une puissance analytique renforcée par un Ensemble AI model

24 *Small Language Models*

Analyses enrichies grâce à la Profiling Base

En moyenne **10 à 100 fois** plus de datapoints intégrés.

Structuration et optimisation du process

Mise à jour de l'évaluation **en temps réel**

Une seule plateforme transversale

La pharmacovigilance, le développement clinique, l'accès au marché, le pipeline manager et les affaires réglementaires.

Une limitation a été observée cependant concernant **la prédiction des bénéfices potentiels,** soulignant ainsi l'importance de projets comme **i-Demo.**

Résultats

Pour la première fois dans l'histoire de Sanofi : nous avons réussi à prédire le rapport bénéfice-risque d'un médicament, en nous appuyant sur un volume de données **9 fois supérieur**, provenant de sources internes et externes.

18 mois de travail réalisés en quelques minutes

$181 million redirigés grâce à la découverte anticipée de 24 risques d'inflammation grave

ArcaScience = $125k VS **CRO** = $1,5m

BR PREDICT : Un développement ambitieux créateur de forte valeur pour ARCASCIENCE

BR Predict : la 1ère plateforme d'analyse B-R prédictive automatisée multimodale pour screener les candidats médicaments les plus prometteurs dès les phases précoces

Notre solution aujourd'hui

- Plateforme opérationnelle éprouvée pour structurer et valoriser massivement des milliards de données clés au B-R, fondation de notre capacité prédictive par IA.

Travaux du projet

- En plus de la profiling base (100b), il s'agira d'Intégrer à la plateforme des modèles de données moléculaires, précliniques, mécanistiques génétiques, impliquant la voie cible, ainsi que des données de réponse thérapeutique en vie réelle, afin de concevoir un World Model spécifique B-R : la première IA prédictive de confiance du domaine.

Impacts projet

- Réduction anticipée du taux d'échec des développements cliniques entre 5 et 10%* grâce à la prédiction du bénéfice-risque précoce.
- Réduction des échecs tardifs et les amendements grâce à l'analyse prédictive multimodale.
- Positionnement de ArcaScience comme le partenaire incontournable du sBR pour démultiplier notre nombre de contrats et assoir durablement la confiance de la filière au complet.

ArcaScience entend devenir la référence mondiale de l'analyse B-R, en accompagnant l'industrie pharmaceutique vers des décisions R&D plus rapides, plus sûres et plus stratégiques.

*optimisation rencontrée en contexte clinique avec la plateforme ArcaScience (~2/5%) + BR Predict

L'évolution de la solution technologique ARCASCIENCE grâce au programme I-Demo

ArcaScience avant I-Demo

Analyse B-R adaptée aux obligations règlementaires

- Développement Clinique
- Accès marché
- Pharmacovigilance
- Gestion de pipeline de candidat
- Affaires réglementaires

Cas d'usage

Maturité cible

ArcaScience après I-Demo (Valeur ajoutée)

Analyse B-R prédictive améliorant le profil clinique amont du candidat

- Recherche et développement préclinique
- Recherche translationnelle
- Business développement

La plateforme conservera ainsi 90 % de composantes génériques, garantissant à la fois adaptabilité et performance et 10% de composantes liées à des pathologies ciblées.

Les bénéfices du programme I-Demo seront démontrés et validés sur le cancer du poumon.

Un choix pertinent pour plusieurs raisons :

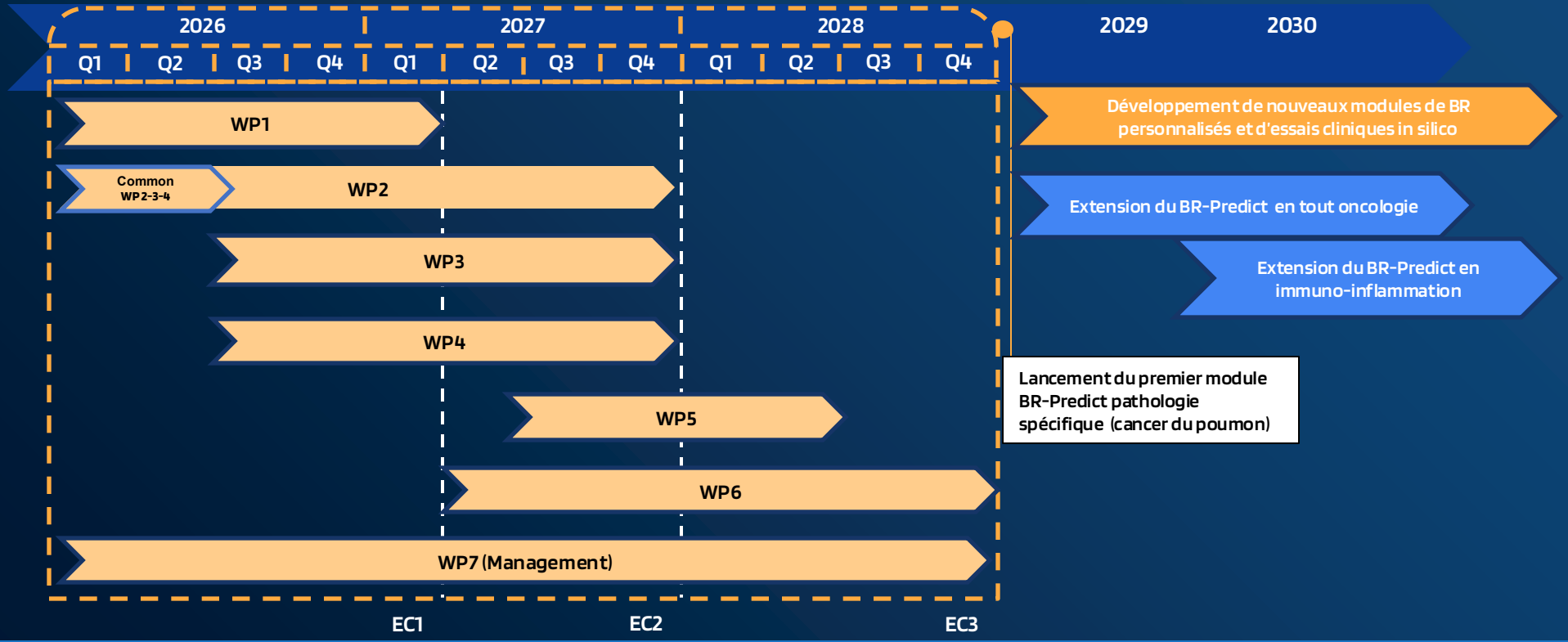
1- Abondance et disponibilité des données : facilite la validation et calibration sur des données fiables et diversifiées.

2- Enjeux contrastés : équilibre entre bénéfice de survie et toxicités variables illustrant la nécessité de ces modèles.

3- Hétérogénéité des profils et thérapies : permet de tester la robustesse du modèle sur des critères multiples (polymorphisme, biomarqueurs, ...)

4- Evolution des stratégies thérapeutiques : constante innovation du développement clinique idéal pour tester l'adaptabilité du modèle prédictif (stratégie émergentes vs standards) à l'intérêt du marché immédiat.

Plan prévisionnel du développement



Structure du Projet : 4 Work Packages "DATA"

Work Package 1

Objectif:
Anticiper les bénéfices et les risques d'une molécule à partir de sa structure chimique seule.

Tâches :

- Qualification des sources de données
- Construction de la base de données
- Calcul des descripteurs
- Modèles QSAR et QSTR (ML)
- Intégration & validation

Résultats Attendus
Prédiction des risques et évaluation des performances
Prédiction des bénéfices et évaluation des performances

Risques identifiés
- Limite d'applicabilité aux structures entièrement nouvelles
- Caractère non exhaustif des structures couvertes (essentiellement biologiques)

Work Package 2

Objectif:
Anticiper les bénéfices et les risques d'une intervention via données in vivo et études toxicologiques.

Tâches :

- Préparation de l'infrastructure (WP 2- 3- 4)
- Qualification des sources de données
- Extraction et structuration des données (NLP)
- Construction de la base de données
- ML pour la prédiction du risque
- ML pour la prédiction de l'efficacité
- Intégration & validation

Résultats Attendus
Modèle NLP pour extraction des données précliniques (noms de modèles animales, caracteristiques, resultats ..)
Prédiction des risques et évaluation des performances
Prédiction des bénéfices et évaluation des performances

Risques identifiés
- Biais de publication et de reporting sélectif
- Limites spécifiques aux biologiques (modèles substitutifs, immunogénicité, faible transposabilité)

Structure du Projet : 4 Work Packages "DATA"

Work Package 3

Objectif:
Anticiper les bénéfices et les risques via les effets sur biomarqueurs pathologiques, cibles moléculaires et polymorphismes génétiques.

Tâches :

- Préparation de l'infrastructure (WP 2- 3- 4)
- Qualification des sources de données
- Extraction et structuration des données (NLP)
- Traitement des données de séquençage
- Construction des bases de données
- ML pour la prédiction du risque
- ML pour la prédiction de l'efficacité
- Intégration & validation

Résultats Attendus
Modèle NLP pour extraction des donnés géntiques (noms d' allèles, séquences, phénotypes ...) et évaluation des performances
Prédiction des risques et évaluation des performances
Prédiction des bénéfices et évaluation des performances

Risques identifiés
- Grande Variabilité inter-individuelle

Work Package 4

Objectif:
Valider relations structure-cibles-bénéfices/risques via données vie réelle (FAERS, EDS, EHDS).

Tâches :

- Préparation de l'infrastructure (WP 2- 3- 4)
- Qualification des sources de données
- Amélioration de l'extraction des données (NLP)
- Construction de la base de données
- ML pour la prédiction du risque
- ML pour la prédiction de l'efficacité
- Intégration & validation

Résultats Attendus
Modèle NLP pour extraction des données cliniques (caractéristiques "patients", caractéristiques "outcomes", caractéristiques "interventions"...) et évaluation des performances
Prédiction des risques et évaluation des performances
Prédiction des bénéfices et évaluation des performances

Risques identifiés
- Qualité et complétude variables des données RWD
- Hétérogénéité entre sources RWD
- **Variabilité inter-pathologies** difficile à modéliser

Structure du Projet : 2 Work Packages "WORLD MODEL"

Work Package 5

Objectif:
Création un cadre de connaissances interopérable intégrant données moléculaires, précliniques, génomiques et réelles dans une couche sémantique unifiée servant d'ontologie fondamentale pour contextualiser et combiner les prédictions des WP1-4.

Tâches :

- Conception et Sélection de l'Ontologie
- Harmonisation des Données et Cartographie des Entités
- Construction du Graphe de Connaissance
- Extraction et Validation des Relations
- Développement interface requête et API

Résultats Attendus
Graphe de connaissances peuplé avec >100K entités et >1M relations
Tables de correspondance validées entre différentes terminologies
API de requête avec documentation

Risques identifiés

- La couverture ontologique peut être incomplète pour les nouvelles cibles ou les maladies rares
- L'exhaustivité du graphe de connaissances dépend de la qualité des données des WP1-4
- Maintenir la cohérence lors de l'arrivée de nouvelles données nécessite une curation continue

Work Package 6

Objectif:
WP6 intègre modèles prédictifs (WP1-4) et paysage de connaissances (WP5) en un "modèle du monde" simulant bénéfice-risque par molécule via relations causales propriétés moléculaires-mécanismes biologiques-caractéristiques patients.

Tâches :

- Conception de l'Architecture du Modèle
- Intégration en Ensemble des WP1-4
- Modélisation causale
- Quantification de l'Incertitude
- Interface de Visualisation
- Validation et Calibration

Résultats Attendus
Premier World Model mondial intégrant des données multimodales pour la prédiction de l'efficacité et des risques des interventions avec évaluation des performances dans les cas spécifiques à une pathologie et non-spécifiques, atteignant une performance de AUC>0.9
Interface de visualisation interactive sous forme de carte mentale

Risques identifiés

- Interactions émergentes
- Données rares pour certains risques

Partenaires du projet

Institutionnels

Institut du Cerveau – ICM – ANS

WP1/4/6 #validation #data

Industriels

sanofi

WP1/2/3/6 #validation

GRADIENT HEALTH

WP4/6 #data

Hospitaliers

Cedars Sinai

WP1/2/4/6 #validation #data

MAYO CLINIC

WP1/3/4/6

#validation #data

En cours de cadrage

AMI Labs

WP6 #validation

Analyse de la concurrence

Le développement de solutions d'analyse de bénéfices-risques automatiques est encore très peu répandu et couvert quasi totalement par des consultants.

ArisGlobal est la solution marché la plus proche de la plateforme ArcaScience. Elle utilise différents modèles IA pour optimiser le processus de l'analyse B/R

Elément différenciant : streamlining de l'évaluation du B-R automatisée sur données mécanistiques precliniques & cliniques, scalables sur plusieurs milliers de candidats à la fois.

Solution de B-R servicielle

- Approches qualitatives ou quantitatives avec maîtrise réglementaire
- Accompagnement personnalisé, mais limités par les nombreuses tâches manuelles.
- Pas d'infrastructure technologique dédiée
- Pas d'industrialisation avec évaluations rapides, répétables et à grande échelle.

YolaRx · BioBoston Consulting · MMS · freyr · DeltaPV · arriello

Solutions AI non B-R

- Capacité AI étendue
- Orientation du développement du candidats médicaments, via analyses mécanistiques, documentaires ou RWE, avec options de safety
- Aucune capacité de réalisation d'évaluations bénéfice-risque ciblées pour un candidat médicament de manière industrialisée.

causaly · nference · IQVIA · OWKIN · QUANTHEALTH · Insilico Medicine · AETION

Une stratégie Go-To-Market permettant de s'étendre rapidement en capitalisant sur les forces existantes de la société

Stratégie déploiement sur la base de clients existant

Add-on BR-Predict avec effet de levier sur l'ensemble de leurs portefeuilles :

1- PoC sur 1 asset
2- Généralisation aire thérapeutique
3- Caractérisation des assets prioritaires
4- Contrats-cadres, facilitant l'utilisation de BR-Predict molécule par molécule et permettant d'industrialiser la collaboration.

Typologie de client : identique.

Nouvelle typologie de client

Ouverture complète de tout le marché drug discovery :

- Big Pharma
- Mid Pharma
- grandes biotechs.

Extension mondiale

Déploiement international immédiat grâce à :

- Soutien des autorités de santé régionales (FDA, EMA, MHRA, HAS)
- Maîtrise des standards et normes locaux via socle ArcaScience, opéré par BR-PREDICT
- Présence réclamée dans tous les congrès internationaux spécialisés safety & B-R
- Une équipe robuste de talents ayant déployé des solutions AI en santé dans le monde, avec succès : notre CMO, CBO, cofondateur notamment

Socle clients:

Modèle économique et pricing de la solution : SaaS privé/molécule

Service PoC

75 k€ - Trial Balancer

150 k€ - Predict

Cas d'usage avant le programme i-demo

Market access

Réglementaire

Pharmacovigilance

Plateforme

=

125 k€ par molécule analysée

Cas d'usage après le programme i-demo

Recherche et développement

Recherche translationnelle et préclinique

Business développement

Module Predict

=

250 k€ par molécule analysée

BR Predict s'appuie sur nos contrats de la plateforme :

1- Revalorisation des pipelines existants (valeur)

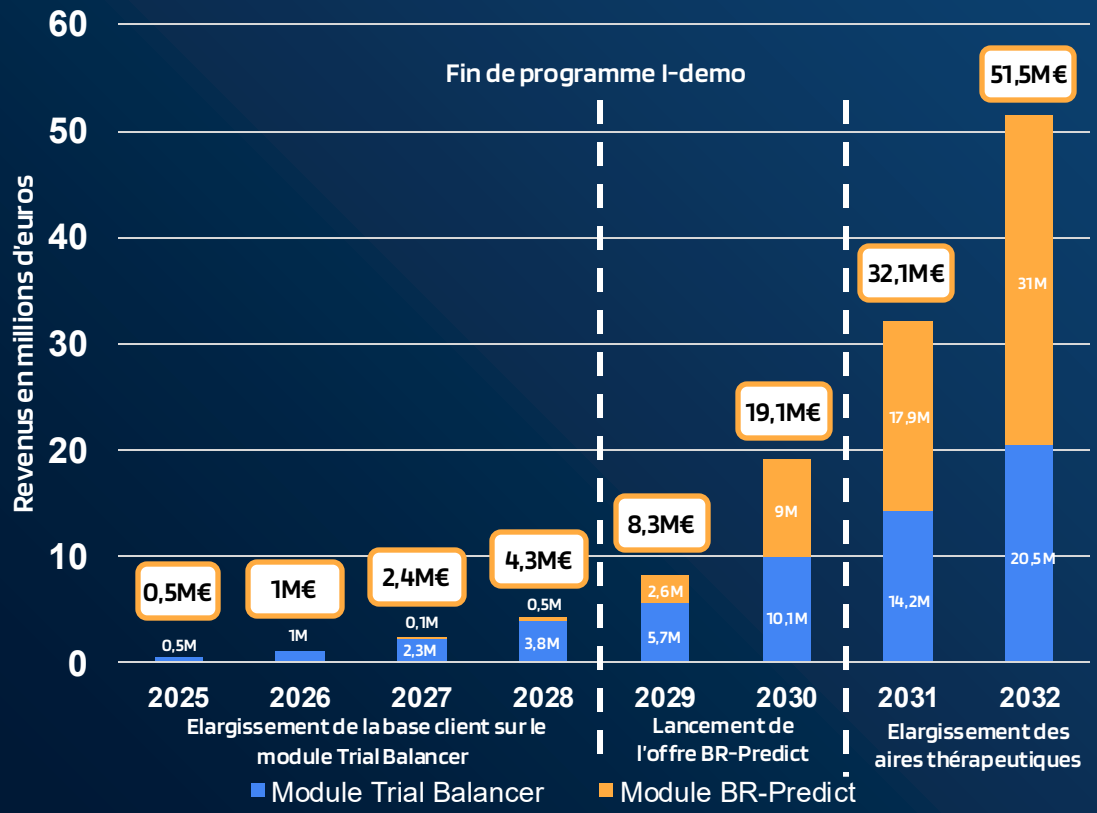
2- Effet de levier sur l'ensemble du portefeuille client (volume)

3- Impact substantiel sur les coûts de développements estimés à **~$7,5m/candidat atteignant la phase 1***

Chaque année, 3k candidats atteignent la phase 1 dans le monde. Cela constitue une opportunité de $22,5b d'économie en R&D.

*FDA Project Optimus a évalué à >50% des coûts avant la phase 1 sont liés à des décisions mal informées, notamment des choix de dose, de design, d'indication, et des signaux de risque safety tardifs

Stratégie de commercialisation et prévisions économiques



Justification des prévisions :

Le module Trail Balancer cible des molécules plus matures, qui sont par nature moins nombreuses dans les portefeuilles R&D. Pour soutenir la croissance, la stratégie repose donc sur un élargissement progressif du portefeuille client, en adressant les grandes biotechs, les mid pharmas et les big pharmas. Le nombre de molécules analysées passe ainsi de 3 en 2025 à 20 en 2028, puis atteint 135 en 2032,

Le module BR-Predict voit une croissance exponentielle auprès des entreprises pharmas déjà clientes. L'offre prévoyant 2 analyses PoC en 2028, 46 molécules analysés auprès des clients Trial Balancer en 2030, puis atteint 199 en 2032. Cette croissance reflète l'intégration progressive du module sur l'ensemble des portefeuilles précliniques, rendue possible grâce à la stratégie de commercialisation menée en parallèle du programme i-Demo et de la très forte polyvalence de la plateforme.

Impacts sociétaux

Impact de l'accélération du développement de médicament

En identifiant plus tôt les candidats les plus prometteurs, la solution accélère l'arrivée de thérapies réellement innovantes pour des populations médicalement délaissées, améliorant concrètement leur accès aux soins et leur qualité de vie.

Réduction du nombre d'individu impliqué dans des essais clinique

Un essai clinique précoce (phase I ou II) mobilise généralement :

- Environ 10 à 50 volontaires sains (en phase I)
- Puis 50 à 300 patients (en phase II), parfois plus selon l'indication

En moyenne, chaque essai annulé ou évité grâce à une meilleure sélection en amont permettrait d'épargner 10 à 300 personnes (volontaires et patients confondus).

Impact de création d'emplois sur le territoire :

Le projet permettra la création de 27 emplois directs, incluant des ingénieurs IA, bioinformaticiens, chefs de projet R&D et profils en affaires réglementaires.
Il prévoit également l'intégration de 4 doctorants et la formation de 8 personnes par an (stagiaires, alternants, jeunes diplômés).
Des emplois indirects seront aussi générés chez les sous-traitants technologiques et scientifiques, renforçant l'écosystème local.

Impact filière

En replaçant l'analyse bénéfice/risque au cœur de la décision, ArcaScience transforme ce qui était perçu comme une contrainte réglementaire en un levier stratégique majeur, offrant aux entreprises pharmaceutiques une nouvelle perspective pour prioriser les candidats les plus prometteurs.

Le projet porte un impact sociétal fort, en contribuant à un développement pharmaceutique plus éthique, plus efficace et plus durable, au bénéfice des patients, des soignants et de l'ensemble du système de santé.

Impacts environnementaux

Réduction des ressources utilisés

La réduction des projets non viables limite la consommation de matières premières, d'eau et de réactifs, permettant une utilisation plus responsable des ressources nécessaires au développement pharmaceutique.

Empreinte carbone

En concentrant les efforts de R&D sur les candidats réellement prometteurs, la technologie diminue les émissions liées au transport, aux synthèses répétées et aux activités cliniques dispersées.

Réduction des déchets

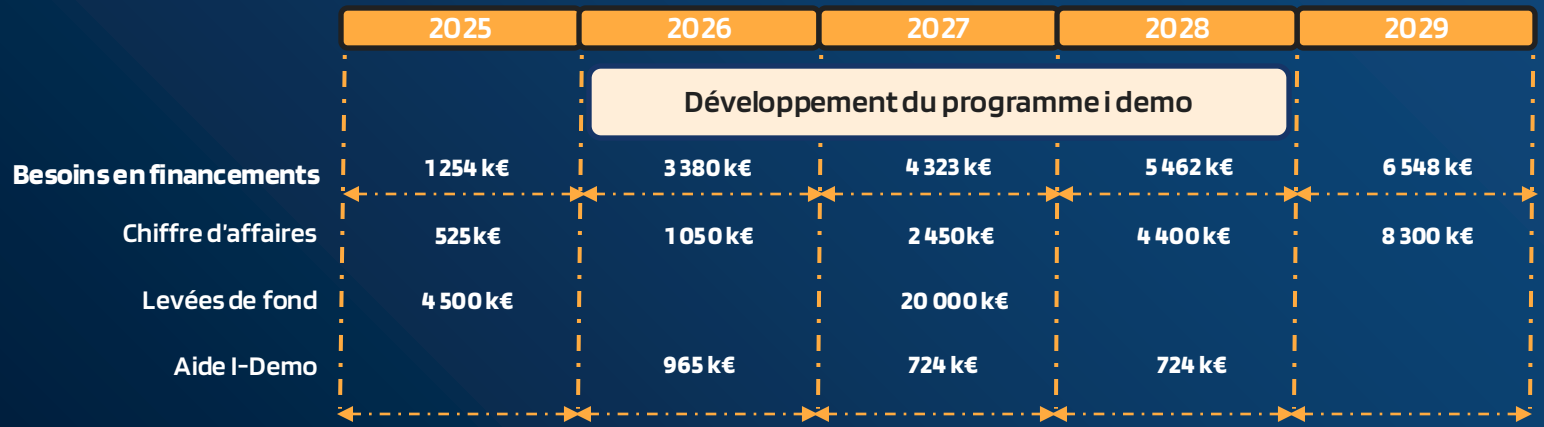
La contraction du nombre d'expérimentations et d'essais évités réduit significativement les déchets chimiques, plastiques et biomédicaux générés dans les laboratoires et les sites cliniques.

Réduction de l'impact environnemental dans le développement de candidats

En limitant le gaspillage scientifique et les projets à forte empreinte écologique, la solution contribue à un modèle de développement pharmaceutique plus durable, plus efficient et plus aligné avec les exigences environnementales actuelles.

Bien que le projet n'ait pas pour objectif principal de générer des impacts environnementaux majeurs, la réduction du nombre d'essais cliniques inutiles, et des ressources qu'ils mobilisent, constitue un levier écologique significatif dans le domaine du développement pharmaceutique.

Plan de financement



- **Levée de fonds 2025 (4,5 M€) : fonds propres sécurisés couvrant le cofinancement complet du projet i-Demo.**
- **Levée prévue Q2-2027 (20 M€) : financement du déploiement commercial de BR-Predict et de l'extension technologique de la plateforme (nouveaux modules, essais cliniques virtuels) grâce à I-Demo.**
- **Soutien i-Demo : accélérateur clé de création de valeur, s'appuyant sur des fonds déjà sécurisés et une traction commerciale initiale pour maximiser le ROI et positionner ArcaScience comme leader du Benefit-Risk.**

Justificatif de l'aide

La technologie ArcaScience, déjà leader en analyse bénéfice-risque, plus précise et plus compétitive que l'état de l'art, doit désormais évoluer pour orienter les décisions de développement clinique en amont.

Le projet i-Demo : un levier stratégique d'innovation

Grâce au soutien du programme i-Demo, dès les phases précliniques et de phase I, nous pourrons:

- Combler les lacunes de la sélection des candidats précoces ;
- Décupler notre potentiel commercial atteignable ;
- Compléter le développement stratégique de la plateforme ARCASCIENCE en devenant l'allié privilégié de la recherche clinique et des régulateurs.

Le projet i-Demo permettra ainsi de positionner ARCASCIENCE comme le premier acteur mondial de l'analyse bénéfice/risque dédiée à la recherche translationnelle et à la stratégie de développement de molécules candidates.

Une ambition industrielle forte, un marché vaste

Contrairement aux solutions existantes, la plateforme ARCASCIENCE adresse :
– un marché large et transverse, couvrant de multiples aires thérapeutiques ;
– l'ensemble du cycle de vie des molécules, de la recherche exploratoire à la commercialisation ;
– un fort potentiel de revenus récurrents, accéléré dès 2028 par l'effet portefeuille des clients.
À ce jour, aucun acteur ne combine une telle couverture avec ce niveau de précision et de flexibilité dans le B-R.

Pourquoi i-Demo est essentiel ?

I-Demo constitue le catalyseur stratégique permettant à ARCASCIENCE d'industrialiser sa technologie, d'accélérer sa vision et de s'imposer comme leader mondial de l'IA appliquée à l'analyse bénéfice/risque. Ce soutien fera émerger un champion national de la pharmatech et soutiendra la création durable de plusieurs dizaines d'emplois hautement qualifiés.

ArcaScience

Merci !

Romain CLEMENT
CEO & Co-Founder

romain@arcascience.ai

+336.38.56.15.21

www.arcascience.ai

Future 4 Care, Biopark, 75013, PARIS

They trust us

Annexe : Présentation du Work Package 1

ArcaScience

Objectif : Le WP1 a pour objectif de prédire les bénéfices et les risques potentiels associés à une molécule en se basant uniquement sur sa structure chimique. La validité de cette approche sera évaluée par le taux de précision obtenu sur des molécules déjà commercialisées.

Tâches :

T1.1. Qualification des sources de données
Identifier, évaluer et qualifier les bases de données pertinentes (NCGC qHTS cytotoxicity, ChEMBL, ToxCast™, etc.)

T1.2. Construction et curation de la base de données
Constituer un dataset consolidé de molécules commercialisées, adjuvants et protéines avec profils bénéfice–risque documentés.

T1.3. Représentation moléculaire & calcul des descripteurs
Transformer les structures chimiques en représentations numériques descriptives et exploitables pour la modélisation

T1.4. Développement de modèles QSAR
Construire des modèles prédictifs d'efficacité thérapeutique à partir de la structure (approches ML plutôt que QSAR classiques).

T1.5. Développement de modèles QSTR
Construire des modèles prédictifs de toxicité et de risques de sécurité (également basés ML).

T1.6. Intégration & validation
Contrôle qualité des données, validation des résultats et évaluation des performances des modèles.

KPI

Seuil de validation : AUC>0.65

Limitations rencontrées :

Une revue de 2025 montre qu'aucun modèle QSAR unique ne peut prédire efficacement l'ensemble de l'espace chimique ; les meilleurs modèles restent performants uniquement au sein de domaines chimiques spécifiques [1].
Les modèles de pointe (DNN, Random Forest, GNN…) offrent d'excellentes performances pour les molécules proches du jeu d'entraînement, mais voient leur précision diminuer sur des structures chimiquement dissemblables [2].
Exemples de performances issues de l'état de l'art :
QSAR pour activité FGFR-1 : $R^2 = 0,7869$ (train), 0,7413 (test) [3]

Livrables :

Modeles de prédiction QSAR et QSTR validés.

Jalon associé :

Être en mesure, à partir d'une structure chimique donnée, de prédire les bénéfices et risques potentiels avec un niveau d'incertitude maîtrisé

Annexe : Présentation du Work Package 2

ArcaScience

Objectif : Le WP2 vise à prédire les bénéfices et les risques potentiels d'une intervention à partir des données in vivo disponibles et des études toxicologiques de la littérature, tout en intégrant un score de fiabilité par pathologie pour chaque modèle préclinique.

Tâches :

T2.1. Préparation de l' infrastructure
Étape commune aux WP2, WP3 et WP4, visant à tester les architectures de modèles et l'infrastructure optimale pour la réalisation de ces workpackages.

T2.2. Qualification des sources de données in vivo
Identification et sélection des études d'efficacité et de toxicité + littérature générale.

T2.3. Extraction et structuration des données (NLP)
Extraction automatique des informations relatives aux modèles in vivo et structuration des données.

T2.4. Création d'une base de données curée et standardisée
Consolidation des données extraites en une base fiable et exploitable.

T2.5. Modèles ML pour la prédiction de l'efficacité
Prédiction du succès clinique à partir des résultats in vivo au niveau de la pathologie / /aspect spécifique de la pathologie,
incluant un score de fiabilité des modèles précliniques.

T2.6. Modèles ML pour la prédiction du risque
Prédiction des risques toxiques humains à partir de données animaux.

T2.7. Validation et intégration
Évaluation des performances, cohérence des données, intégration dans les pipelines internes.

KPI

KPI principal : précision des prédictions pour des molécules déjà commercialisées, supérieure à celle obtenue dans le WP1

Limitations rencontrées :

Données propriétaires inaccessibles, contenus sous paywall. - Biais de publication et de reporting sélectif→ augmentation potentielle du risque d'erreur, Limites spécifiques aux biologics (modèles substitutifs, immunogénicité, faible transposabilité).

Livrables :

Modèle NLP fonctionnel pour extraction d'informations in vivo.
Score de fiabilité par modèle préclinique et par pathologie/aspect.
Modèles ML prédictifs des bénéfices et risques basés sur données in vivo.
Capacité finale : prédire les bénéfices et risques potentiels pour une structure donnée.

Milestone associé :

Définition d'objectifs de précision selon le périmètre (spécifique / non spécifique).
Exemple (état de l'art) : des modèles hybrides ML + mécanistiques pour la prédiction PK in vivo ont atteint 64 % (clearance) et 62 % (volume de distribution) de prédictions dans une marge d'erreur ×2.

Annexe : Présentation du Work Package 3

ArcaScience

Objectif : Le WP3 vise à prédire les bénéfices et risques d'une molécule à partir de ses effets sur les biomarqueurs de la pathologie, de sa cible moléculaire et des polymorphismes génétiques associés. La prédiction est évaluée par sa précision sur des molécules commercialisées.

Tâches :

T3.1. Qualification des sources
- Biomarqueurs/pathologie
- Variants génétiques des cibles

T3.2. Extraction et structuration des données (NLP)
- Extraction automatique des polymorphismes génétiques
- Structuration des effets fonctionnels

T3.3. Construction d'une base de données intégrée
- Molécule → Biomarqueurs/génomique → Pathologie
- Molécule → Cible → variants génétiques → impacts fonctionnels

T3.4. Modèle ML – Bénéfice
- Input : biomarqueurs/génomique de la pathologie + Cible
- Output : prédiction de bénéfice

T3.5. Modèle ML – Risque
- Input : génotype de la target + mécanisme d'action du traitement
- Output : prédiction de risque

T3.6. Validation & Intégration
- Évaluation des performances
- Harmonisation et intégration dans les pipelines internes

KPI :

Cette prédiction est évaluée par sa précision sur des molécules déjà commercialisées.

Raison d'être du WP :

- Approche basée sur la médecine personnalisée.
- Applicable à tout type de traitement et toute pathologie.
- Extension avancée du QSIIR.
- De nombreuses études démontrent l'impact du profil génétique sur l'efficacité thérapeutique, notamment en oncologie.
- Cette approche n'a pas encore été testée à grande échelle, ce qui en fait un axe d'innovation fort.

Livrables :

- Modèle NLP fonctionnel pour extraction des informations génétiques
- Base de données : cibles ↩ biomarqueurs ↩ pathologie ↩ variants génétiques
- Modèles ML prédictifs : bénéfices & risques basés sur la génétique
- Capacité finale : prédire bénéfices et risques pour une structure donnée via génomique + cible

Milestone associé :

Définition des objectifs de précision pour un modèle généraliste, pour prédiction par cible (cible générale non pathologie spécifique)

Exemples de performances de l'état de l'art :
- UGenome (2025) : modèles pharmacogénomiques IA jusqu'à 99 % de précision pour prédire les réponses thérapeutiques via données multi-omiques, validés par comparaison aux essais cliniques et recommandations de régulateurs (ex. FDA) [1].
- Park et al. (2023) : modèles ML/DL prédictifs de réponse (lnIC50) basés sur profils d'expression et mutations dans divers cancers ; résultats robustes mais spécifiques à chaque traitement et contexte cellulaire [2].
- Miranda et al. (2021) : modèles intégrant polymorphismes + variables cliniques pour prédire réponse et EIs au tamoxifène ; performances encourageantes mais contexte très spécifique [3].

Annexe : Présentation du Work Package 4

ArcaScience

Objectif : L'extraction des données patients ainsi que des données de sécurité et d'efficacité issues de bases de vie réelle (FAERS, EDS, EHDS…) permet de valider et renforcer les relations identifiées entre structure moléculaire, modèles précliniques, cible thérapeutique, bénéfices et risques.
Le WP4 vise à développer des modèles prédictifs de bénéfices et de risques basés sur les données de vie réelle, intégrant :

- **les conditions d'utilisation en pratique,**
- **les populations non sélectionnées,**
- **les comorbidités, la polymédication,**
- **les facteurs socio-démographiques.**

Tâches :

T4.1. Qualification des sources RWE
Identifier, évaluer et qualifier les bases FAERS, EDS, EHDS et autres sources utilisables.
T4.2. Extraction & structuration des données RWE
Extraction, harmonisation et standardisation des données de sécurité et d'efficacité.
T4.3. Modèles prédictifs RWE – Bénéfices
Entraînement de modèles ML prédictifs du bénéfice clinique en conditions réelles.
T4.4. Modèles prédictifs RWE – Risques
Entraînement de modèles ML prédictifs des risques/ADRs en vie réelle.
T4.5. Validation & calibration
Calibration externe, évaluation robuste (AUC, accuracy, calibration plots), comparaison aux modèles WP1–3.

KPI :

précision des modèles RWE supérieure à celle obtenue via WP1/2/3.

Livrables :

- Modèles ML prédictifs basés sur données RWE (bénéfices et risques).
- Pipeline complet d'extraction, standardisation et analyse des données RWE.
- Modèle généralisable intégrant facteurs cliniques réels + données moléculaires + données de pathologie (WP1–3 + WP4).

Milestone associé :

un milestone réaliste pour un modèle généralisé, tenant compte de la variabilité inter-pathologies, des biais RWE et de l'hétérogénéité des populations.
Un milestone plus élevés pour le cancer de poumon

Annexe : Présentation du Work Package 5

ArcaScience

Objectifs : Le WP5 vise à créer un cadre de connaissances structuré et interopérable qui intègre de multiples sources de données hétérogènes (structures moléculaires, résultats précliniques, données génomiques, données de vie réelle) dans une couche sémantique unifiée. Ce « paysage de connaissances » sert d'ontologie fondamentale et d'architecture de données permettant de contextualiser, comparer et combiner les prédictions des WP1-4.

Tâches :

T5.1. Conception et Sélection de l'Ontologie
Identifier et adopter des ontologies standard (MedDRA, SNOMED CT, ChEBI, Disease Ontology) et définir les relations entre entités moléculaires, biologiques et cliniques

T5.2. Harmonisation des Données et Cartographie des Entités
Associer les résultats des WP1-4 à des terminologies standardisées et créer des références croisées entre sources de données et résultats

T5.3. Construction du Graphe de Connaissances
Construire la structure de la base de données graphe (nœuds = entités, arêtes = relations) et la peupler avec les données intégrées des WP1-4

T5.4. Extraction et Validation des Relations
Utiliser le traitement automatique du langage naturel pour extraire des relations supplémentaires de la littérature, valider les relations par rapport aux associations connues médicament-maladie-résultat, et attribuer des scores de confiance aux relations)

T5.5. Développement de l'Interface de Requête et de l'API
Créer un langage de requête pour les questions complexes multi-niveaux, développer une API permettant au WP6 d'accéder au graphe de connaissances, et implémenter la mise en cache pour les requêtes fréquentes

Raison d'être du WP :

Les WP1-4 génèrent des prédictions isolément en utilisant différents types de données et terminologies (descripteurs chimiques, critères d'évaluation précliniques, variants génétiques, résultats de données de vie réelle)

Sans standardisation, ces prédictions ne peuvent pas être comparées ou intégrées de manière significative

Un graphe de connaissances permet :

- Interopérabilité sémantique : Associer « hépatotoxicité » dans les études précliniques à « élévation des ALAT » dans les données de vie réelle et « lésion hépatique » dans les notices cliniques
- Découverte de relations : Relier cibles moléculaires → voies métaboliques → maladies → résultats
- Propagation de l'incertitude : Tracer la provenance des données et le niveau de confiance entre les sources
- Explicabilité : Permettre la traçabilité de la prédiction jusqu'aux données probantes sources

KPI :

Intégration réussie de ≥90% des entités de données des WP1-4 dans un graphe de connaissances standardisé avec des relations sémantiques validées. Temps de réponse aux requêtes<2 secondes pour les requêtes complexes multi-sources.

Limitations rencontrées :

- La couverture ontologique peut être incomplète pour les nouvelles cibles ou les maladies rares
- L'exhaustivité du graphe de connaissances dépend de la qualité des données des WP1-4
- Maintenir la cohérence lors de l'arrivée de nouvelles données nécessite une curation continue

Livrables :

- Graphe de connaissances peuplé avec >100K entités et >1M relations
- Tables de correspondance validées entre différentes terminologies
- API de requête avec documentation
- Rapport de qualité des données montrant la couverture et les scores de confiance

Milestone associé :

- Schéma ontologique finalisé et validé avec les experts du domaine
- Graphe de connaissances peuplé avec les données des WP1-3 (couverture à 80%)
- Intégration complète des données de vie réelle du WP4 et validation des relations terminée
- API de requête opérationnelle et testée selon les exigences du WP6

Annexe : Présentation du Work Package 6

ArcaScience

Objectifs : Le WP6 intègre tous les modèles prédictifs (WP1-4) avec le paysage de connaissances (WP5) pour créer un « modèle du monde » capable de simuler des profils bénéfice-risque pour n'importe quelle molécule. Ce modèle prédit les résultats au niveau individuel, les interactions médicamenteuses et les risques spécifiques au contexte en apprenant les relations causales entre les propriétés moléculaires, les mécanismes biologiques et les caractéristiques des patients.

Tâches :

T6.1. Conception de l'Architecture du Modèle

Concevoir une architecture de réseau neuronal pour intégrer les prédictions des WP1-4, implémenter des mécanismes d'attention pour pondérer les preuves de différentes sources, et construire un module de dynamique temporelle pour les prédictions dépendantes du temps

T6.2. Intégration en Ensemble des WP1-4

Combiner les prédictions en utilisant des pondérations apprises (pas une simple moyenne), gérer les contradictions (par ex., WP2 prédit une toxicité, WP4 données de vie réelle montrent une sécurité), et implémenter la résolution de conflits avec propagation de l'incertitude

T6.3. Modélisation Causale et Prédiction des Interactions

Entraîner des modèles pour prédire les interactions médicament-médicament en utilisant les données moléculaires + données de vie réelle, modéliser les interactions gène-médicament pour les prédictions pharmacogénomiques, et simuler les relations dose-réponse et temps jusqu'à l'événement

T6.4. Quantification de l'Incertitude

Implémenter des approches bayésiennes ou d'ensemble pour les intervalles de confiance, signaler les prédictions avec faible confiance en raison de données éparses, et fournir des explications interprétables des sources d'incertitude

T6.5. Interface de Visualisation (Carte Mentale)

Visualisation interactive du réseau bénéfice-risque (molécule → cibles → voies métaboliques → résultats), permettre l'exploration approfondie du résumé de haut niveau aux preuves sources, et afficher les scores de confiance et la provenance des données

T6.6. Validation et Calibration

Validation rétrospective sur les médicaments retirés du marché (le modèle peut-il prédire les échecs ?), suivi prospectif des prédictions pour les molécules en développement, et graphiques de calibration pour s'assurer que les probabilités prédites correspondent aux fréquences observées

Raison d'être du WP

L'évaluation traditionnelle du rapport bénéfice-risque est :

- Statique (instantané à un moment donné)
- Au niveau populationnel (non personnalisée)
- Manuelle (nécessite une synthèse humaine de données disparates)

Un modèle du monde permet :

- Simulation dynamique : Prédire comment le rapport bénéfice-risque évolue, avec des modifications de dose, ou dans différents sous-groupes de patients
- Tests de scénarios : Simuler « et si nous dosions différemment ? » ou « et si le patient avait ce génotype ? » sans essais cliniques réels
- Quantification de l'incertitude : Intervalles de confiance explicites sur les prédictions, identifiant les lacunes de données
- Raisonnement causal : Aller au-delà de la corrélation pour comprendre les relations mécanistiques (par ex., « Inhibition de la cible → perturbation de la voie métabolique → événement indésirable »)

Cela représente un passage d'un rapport bénéfice-risque descriptif (qu'est-ce qui s'est passé ?) à prédictif (que va-t-il se passer ?) à prescriptif (que devons-nous faire ?).

KPI

- Sensibilité et spécificité élevées pour la prédiction d'événements indésirables graves sur des molécules de test
- Capacité à générer des profils bénéfice-risque cohérents incluant ≥3 sources de données indépendantes
- Validation utilisateur : ≥70% des experts évaluateurs jugent les prédictions comme « cliniquement plausibles »

Limitations rencontrées :

- Rareté des données : Les nouveaux squelettes moléculaires ou les maladies rares peuvent avoir des données d'entraînement insuffisantes
- Interactions émergentes : Impossibilité de prédire des mécanismes complètement nouveaux absents des données d'entraînement

Livrables :

- Modèle du monde intégré combinant les résultats des WP1-5
- Module de prédiction des interactions
- Interface de visualisation interactive sous forme de carte mentale
- Rapport de validation sur 200+ molécules commercialisées et retirées du marché
- Documentation utilisateur et supports de formation
- Système prêt au déploiement avec quantification de l'incertitude

Milestone associé :

- Démonstration spécifique à une maladie : Sélectionner un domaine thérapeutique (cancer du poumon) et démontrer le flux de travail complet :
- Entrée : Nouvelle molécule en Phase 1-2
- Sortie : Profil bénéfice-risque complet avec :
  - Efficacité prédite (ensemble WP1-4)
  - Profil de sécurité prédit (toxicités organiques, événements indésirables)
  - Comparaison aux thérapies existantes
  - Estimations d'incertitude
  - Visualisation au format carte mentale
- Validation : Examen par un panel d'experts pour la plausibilité clinique
- Système généralisé : Modèle du monde opérationnel applicable à plusieurs domaines thérapeutiques, avec :
  - AUC >0,8 pour la prédiction d'événements indésirables graves (ensemble de validation externe)
  - Taux de faux négatifs <10% pour les interactions médicament-médicament connues
  - Tests d'acceptation utilisateur avec ≥5 partenaires pharmaceutiques
- Documentation conforme à la réglementation démontrant la conformité aux lignes directrices sur l'IA

Un conseil scientifique, technique et stratégique de premier plan

Laurent Romary, PhD : Directeur de recherche émérite INRIA, président de DARIAH (infrastructure de recherche numérique pour les arts et les sciences humaines) ;

Philippe Peyre, PhD : Senior VP et secrétaire général chez Sanofi

Kerry Coffee : Fondatrice de Schiller Discovery Healthtech Venture

Geneviève Laurans : Administratrice et CFO de Quotium Technologies

Professeur Clemens M. Schirmer, MD, PhD : Neurochirurgien et professeur à la Geisinger School of Medicine

Hal Lavender : AI, Data and Business Transformation Leader VP chez IBM

Béatrice du Sordet : Fondatrice d'Aplusa (étude de marché de santé)

Professeur Alexis Brice, MD, PU-PH : Président fondateur de l'Institut du Cerveau (ICM), membre de l'Académie des Sciences

Serge Bauin, PhD : Ancien directeur général délégué à la science au CNRS et expert Data Science ministériel

Jean Lopategui, MD, PhD : Directeur du département Pathology de Cedar-Sinaï, Los Angeles, Immuno-hématologue

Fabien Lanteri : Directeur stratégie santé et innovation au sein d'EIT Health (programme d'apprentissage en santé) et expert au Génopole

Emmanuel Capitaine, MD : Coordinateur médical international chez Vivalto Santé (groupe de cliniques privées), ex-responsable innovation chez AstraZeneca

Jordan Lazaro-Gustave : Fondateur d'AAVE, protocole de finance décentralisée qui permet aux utilisateurs de prêter et d'emprunter des cryptomonnaies).

Stéphane Rouault, PhD : Clinical Innovation Lead chez Argenx (société belgo-néerlandaise spécialisée dans l'immunologie) et Roche

Un écosystème de partenaires étendu représentant un soutien de taille

Partenaires de recherche et académiques

Inria — Institut du Cerveau ICM

ArcaScience collabore avec l'ICM et l'Inria sur des projets de recherche avancés, notamment pour identifier de nouveaux candidats thérapeutiques et contribuer à un projet européen IHI sur le cancer du cerveau pédiatrique.

Ecosystème et structure de soutien

Cedars Sinai — PLUGANDPLAY — Business France — Future4care

Sa présence au sein de Future4Care lui donne accès à un écosystème riche réunissant startups, groupes pharmaceutiques, institutions et investisseurs, renforcé par l'accompagnement de Business France depuis 2025 et son intégration dans la communauté international Plug and Play et Cedars Sinai.

Partenaires commerciaux

ICON — Mapi Research Trust — sanofi — VIDAL — AstraZeneca — BIOCODEX — DILON TECHNOLOGIES

ArcaScience travaille avec les principaux acteurs engagés dans la recherche et le développement de nouveaux médicaments. Ses solutions s'adressent aux grandes entreprises pharmaceutiques, aux sociétés de biotechnologie de toutes tailles ainsi qu'aux CRO qui conduisent une part majeure des essais cliniques pour le secteur.

À la fin juin 2025, ArcaScience compte 9 contrats signés avec 7 clients, parmi lesquels des leaders internationaux tels que Sanofi, AstraZeneca, Vidal et Biocodex. L'entreprise collabore également avec ICON, y compris sa filiale MAPI, l'une des plus grandes CRO mondiales, ainsi qu'avec des acteurs spécialisés comme Dilon Technologies.