**Cahier des charges pour étudier les performances du recueil de données d’accès précoce au sein du projet « Données d'Accès précoces Normalisées extraites par Traitement automatisé du langage en vue de leur Évaluation » (DANTE) : étude comparative entre une saisie manuelle et une extraction via une intelligence artificielle**

# Notre démarche

La Filière Intelligence Artificielle & Cancers est une association à but non lucratif issue du Contrat Stratégique de Filière des Industries et Technologies de Santé (CSF-ITS) créée le 3 août 2021.

Elle fédère des membres publics (l’Institut national du cancer, l’Agence d’Innovation en Sante et le Health Data Hub) et privés (l’Alliance pour la Recherche et l’Innovation des Industries de Santé, France Biotech et 9 des plus grands industriels de santé mondiaux en oncologie) au sein d’une initiative nationale qui vise à rendre attractif l’écosystème d’innovation en oncologie.

Son ambition est de valoriser l’écosystème national en menant des projets preuves de concept en oncologie qui sont innovants et emblématiques des questions des industriels de la santé.

Afin de mener à bien ces projets, la Filière propose des cahiers des charges pour décrire des prestations attendues autour de ces projets et stimuler l’écosystème d’innovation sur des cas réels.

Ces cahiers des charges sont publics. Les candidats qui confirment leur intérêt à répondre au cahier des charges acceptent alors de s’engager dans une démarche de dialogue compétitif qui vise à retenir un lauréat financé par la Filière.

Les éléments ci-après décrivent le projet DANTE et le cahier des charges des prestations associées à ce projet.

# Contexte et rationnel du projet

Depuis la réforme de l’accès dérogatoire aux médicaments le 1er juillet 2021, l’octroi d’une autorisation d'accès précoce (AP) par les autorités de santé implique un recueil de données en vie réelle chez les patients. Les données à collecter sont définies dans un protocole d’utilisation thérapeutique et de recueil des données (PUT-RD) validé par la Haute Autorité de santé (HAS).

Le projet proposé démarre d’un constat : si les données recueillies dans ce cadre sont d’un grand intérêt – *elles permettent notamment de documenter les premières utilisations de traitements innovants en dehors des essais cliniques et de fournir de précieuses données sur l’usage en vie réelle de ces traitements* – leur collecte se heurte à plusieurs défis : finalité du recueil des données d’accès précoce pas toujours très « claire » pour les professionnels de santé, manque de ressources humaines dans les établissements de santé, redondance avec les données cliniques déjà saisies en routine dans les dossiers patient informatisés (DPI), hétérogénéité des plateformes de saisie des données pour chaque protocole d’utilisation thérapeutique et de recueil des données.

Ces contraintes freinent l’adhésion des professionnels de santé, entrainent des lacunes dans la qualité et la cohérence des données collectées, avec un taux de complétion des données bien inférieur à l’objectif de 90% fixé par les autorités de santé.

Les progrès récents en IA offrent des opportunités prometteuses pour améliorer les processus de saisie et de gestion des données. Les outils basés sur le *natural language processing* (NLP) et le *machine learning*, ont le potentiel d’accélérer la saisie des données, de réduire les erreurs humaines et d’améliorer la qualité des informations collectées, tout en réduisant le fardeau lié au recueil des données pour les professionnels et les établissements de santé.

**L’objectif de ce projet pilote est de mener une étude pour évaluer et comparer les performances des deux méthodes de recueil de données (manuelle « traditionnelle » et par IA) et évaluer leur capacité à répondre aux exigences de qualité, fiabilité et cohérence des informations collectées nécessaires dans le cadre des dispositifs d’accès précoce.**

# Description de la conduite du projet

**Ce projet pilote est un projet de filière, multipartite,** qui engage plusieurs industriels pharmaceutiques, chacun menant son propre accès précoce de manière indépendante, selon son propre calendrier de mise en œuvre.

À ce jour, trois industriels (AstraZeneca, MSD, et Amgen) ont confirmé leur participation au projet, dont deux avec des AP qui pourraient démarrer début 2026, sous réserve d’octroi de l’autorisation par la HAS. D’autres laboratoires pourront rejoindre le projet au fil de l’eau, augmentant ainsi la diversité et le volume des données disponibles pour mener l’étude comparative.

Chaque industriel sélectionnera la CRO de son choix pour l’accompagner dans la gestion « habituelle » de son AP et pour coordonner la collecte des données de façon manuelle. Ces CRO, dites « primaires » dans le présent projet, collaboreront ainsi avec les établissements de santé demandant accès au traitement en AP pour assurer la saisie manuelle des données, conformément aux PUT-RD préalablement validés par la HAS et sans intervention dans le cadre de ce projet.

Simultanément, la Filière IA & Cancers vérifiera, à chaque demande d’accès au traitement validée par le laboratoire, si l’établissement du médecin prescripteur dispose d’une solution d’IA **installée et fonctionnelle** (en propre ou via un prestataire). Si tel est le cas, par la Filière proposera une convention à l’établissement de santé ou à son prestataire IA, **dits « partenaires technologiques »**, en vue d’extraire les données des patients inclus dans cet établissement.

Enfin, au terme de l’étude, une **base de référence** comprenant les données des patients inclus devra être constituée, par exemple par des attachés de recherche clinique (ARC) à partir des DPI. Ce triple recueil (manuel, IA, référence) permettra d’évaluer et comparer, par rapport à une base de référence, les performances des recueils manuel et IA, en termes d’exactitude, d’erreur, de complétude et de concordance des données.

**Le prestataire recherché** **dans le cadre du présent cahier des charges, appellé ci-après « CRO centrale », sera chargé de centraliser et fusionner les données manuelles, transmises par les CRO primaires, et les données IA, transmises par les partenaires technologiques ; de constituer la base de référence dans les centres participant ; apparier les patients des trois bases de données et réaliser les analyses comparatives. Le détail est fourni au paragraphe suivant.**

**L’objectif est d’inclure 147 patients dans l’étude**. Le nombre de centres participant n’est pas connu a priori. Tous les centres prescripteurs disposant d’une solution d’IA (en propre ou via un prestataire) permettant d’extraire les données du set prédéfini de variables à partir des DPI pourront participer. Les centres seront ouverts au fil des inclusions et jusqu’à atteinte de l’objectif d’inclusion de patients.

**Il est précisé que l’étude comparative ne portera pas sur l’intégralité des variables des PUT-RD mais sur un set d’une vingtaine de variables communes (voir Annexe 2)**.Ce set de variables à collecter, défini à partir du modèle de PUT-RD de la HAS[[1]](#footnote-2), sera commun à tous les accès précoces de ce projet, indépendamment de l’indication thérapeutique, et comprendra des informations issues des fiches de demande d’accès au traitement, d’instauration, de suivi et d’arrêt de traitement.

Ces variables incluront des typologies de données supposées « simples à collecter » (données directes) et d’autres plus complexes (données composites, données nécessitant un traitement ou une interprétation approfondie).

# Description détaillée des prestations attendues

Nous recherchons un prestataire capable de :

1. Agir en tant que **tiers de confiance** dans ce projet multipartite et complexe.
2. **Recevoir, assurer la** **traçabilité et stocker de façon sécurisée** et conformément aux règlementations en vigueur, notamment le RGPD, l’ensemble des jeux de données pseudonymisées provenant de chacun des bras de l’étude (*n* CROs primaires et *n* partenaires technologiques)
3. À la fin du suivi des patients inclus (à la fin des AP participant au projet), mais également lors d’un point intermédiaire (Q4 2026, cf. Calendrier prévisionnel de l’étude au § 6 de ce document) et pour extraire des premiers résultats, **constituer une base de données de référence composée des données des patients inclus à partir de leur DPI.** Cette activité pourra comprendre le développement d’un masque de saisie correspondant au set prédéfini de variables, et pourra comprendre l’envoi d’ARC dans chaque centre participant. La Filière reste ouverte à d’autres méthodes, innovantes, qui devront garantir la robustesse de la base de référence, fondamentale pour l’étude.
4. **Fusionner et apparier les données transmises** par les CRO primaires, celles transmises par les partenaires technologiques, et celles de la base de données de référence.
5. **Assurer le data management** des différentes bases de données, incluant les modalités de validation et de nettoyage des données, les critères de qualité des données et les étapes préalables à l’analyse statistique.
6. **Rédiger un plan d’analyse statistique** pour mener l’étude comparative décrite dans le **protocole d’étude - déjà rédigé par la Filière IA & Cancers et dont un extrait et fourni en Annexe 2, ci-dessous page 9**, en collaboration avec les équipes industriels partenaires, un comité d’experts, la HAS et la CNIL. Il est important de noter que les analyses ne se limiteront pas à des indicateurs macroscopiques, mais analyseront les résultats par variables, car toutes n’ont pas la même difficulté de structuration ni le même poids informatif (**cf. Annexe 1 en page 8**). Tout candidat intéressé peut se procurer le protocole d’étude d’ensemble en écrivant à l’adresse suivante : lucas.elle@filiere-ia.fr
7. **Effectuer les analyses statistiques** conformément au plan d’analyse statistique,
8. **Rédiger** **un rapport d’analyse statistique final** intégrant l’ensemble des résultats de l’étude (figures et tableaux inclus). Le rapport d’étude final sera rédigé par la Filière IA & Cancers.

Voici un schéma d’ensemble des flux de données de la prestation demandée (ci-dessous en vert, les actions et données en charge du prestataire) :



Figure - schéma d’ensemble des flux de données la prestation demandée

# Budget de l’étude

Le budget est fixé à **100 000 € hors taxes (cent mille euros hors taxes)**.

Il constitue l’enveloppe financière pour la réalisation de l’ensemble des prestations décrites dans le présent cahier des charges. Toute proposition devra démontrer la capacité du candidat à garantir un haut niveau de qualité scientifique et opérationnelle.

# Calendrier prévisionnel de l’étude

**Q4 2025** : dépôt des premières demandes d’autorisation d’accès précoce auprès de la HAS par les laboratoires participant à DANTE avec un AP.

**Q1 2026** : octroi de l’autorisation d’AP et démarrage des AP avec mise à disposition du traitement et premières demandes d’accès.

**Q4 2026** : première revue de la concordance des données.

**2027 ou 2028** : à la fin des AP, constitution de la base de référence et analyses finales.

# Critères de sélection des partenaires

Les propositions seront évaluées sur la base des cinq critères suivants :

* Compréhension du projet, de ses objectifs et de sa portée.
* Capacité à garantir la sécurité et la conformité des traitement de données.
* Réalisme et pertinence de proposition innovantes – en particulier sur la constitution de la « base de données de référence ».
* Compétences en analyses statistiques.
* Force de propositions pour faciliter la mise en œuvre et le suivi du projet.

# Modalités de candidature et calendrier du dialogue compétitif

Voici une vision d’ensemble du déroulé du dialogue compétitif, entièrement sur la fin de l’année 2025 :



**La procédure de sélection se déroulera en deux phases successives :**

1. **Du 13 octobre au 14 novembre**, la phase de « diffusion d’un appel à manifestation d’intérêt et réception des candidatures ». Les candidats intéressés devront transmettre, **avant le 7 novembre minuit**, une **réponse synthétique** en **3 à 5 pages maximum**. Cette réponse aura pour objet de :
* **Présenter (brièvement) l’organisation et ses compétences clés**, ainsi que l’identification des **personnes pressenties** pour piloter l’étude.
* Proposer une **méthodologie de mise en œuvre et suivi concise**, démontrant la compréhension du projet et des prestations attendues.
* Mettre en avant des **propositions innovantes pour la réalisation de la prestation**, dans le respect de ses finalités.
* Mettre en avant, le cas échéant, les **éléments de différenciation** (ex. expertise en data management, expérience en analyses statistiques).
1. À l’issue de cette phase, **deux à trois candidats** **seront présélectionnés** pour participer à la phase de dialogue compétitif, **du 17 novembre au 19 décembre**. Les candidats présélectionnés seront alors conviés à élaborer des offres détaillées, **à rendre le 5 décembre minuit au plus tard**, pour initier ensuite un dialogueen plusieurs « rounds »avec l’équipe projet de la Filière IA & Cancers et ses partenaires industriels. Cette phase de dialogue visera à :
* **Approfondir la méthodologie proposée** et sa mise en œuvre concrète. En d’autres termes, ajuster et clarifier, par un dialogue direct avec les équipes, les innovations proposées et les propositions techniques et financières.
* **Discuter des modalités pratiques** (gestion des données, envoi d’ARC, respect des réglementations) pour cibler une prestation innovante et réaliste

À l’issue de ce dialogue, **le 19 décembre, un lauréat sera retenu** et mandaté pour la réalisation de la mission.

1. **Statut du lauréat en terme de responsabilité de traitement**

Le lauréat retenu dans le cadre du présent appel à candidatures interviendra en qualité de **prestataire responsable de la mise en œuvre des analyses pour le responsable de traitement**, **la** **Filière IA & Cancers**.

Le lauréat devra **apporter la preuve de son expérience et de son savoir-faire** dans la réalisation de missions comparables, en particulier en matière de gestion, sécurisation, traitement et analyse de données de santé à caractère personnel. Cette démonstration pourra être effectuée au travers de références, d’exemples de missions similaires, et de tout élément justificatif attestant de la qualité et de la conformité de ses pratiques.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Annexe 1 : set de données à collecter

| **Fiche du PUT-RD**  | **Catégorie de la fiche du PUT-RD**  | **N° et description de la variable** | **Valeurs attendues** |
| --- | --- | --- | --- |
| NA | NA | 1 | Médicament en AP | Liste prédéfinie selon les laboratoires participant |
| Demande d'accès | Identification du patient | 2 | Date de naissance (MM/AAAA) | mm/aaaa |
| Identification du patient | 3 | Sexe | * Masculin
* Féminin
 |
| Instauration du traitement  | Conditions d'utilisation | 4 | Date de 1ère administration ou d'instauration du traitement | jj/mm/aaaa |
| Posologie prescrite | 5 | Posologie 1 : dose | Numérique |
| 6 | Posologie 2 : dose | Numérique |
| 7 | Posologie 1 : voie d'administration | * Orale
* Intraveineuse
* Sous-cutanée
 |
| 8 | Posologie 2 : voie d'administration | * Orale
* Intraveineuse
* Sous-cutanée
 |
| Effet(s) indésirable(s) / Situation(s) particulière(s) | 9 | Y a-t-il eu apparition d’effet(s) indésirable(s) ou une situation particulière ? | * Oui
* Non
 |
| Suivi(s) du traitement (autant d'occurrences que d'administrations du traitement) | Suivi du traitement  | 10 | Date de la visite de suivi | jj/mm/aaaa |
| Posologie prescrite | 11 | Posologie 1 : dose | Numérique |
| 12 | Posologie 2 : dose | Numérique |
| 14 | Posologie 1 : voie d'administration | * Orale
* Intraveineuse
* Sous-cutanée
 |
|  | 15 | Posologie 2 : voie d'administration | * Orale
* Intraveineuse
* Sous-cutanée
 |
| Interruption/arrêt temporaire | 16 | Date d'interruption | jj/mm/aaaa |
| 17 | Motif d'interruption | * Progression de la maladie
* Effet indésirable
* Souhait du patient
* Autre (préciser)
 |
| 18 | Date de reprise | jj/mm/aaaa |
| Effet(s) indésirable(s) / Situation(s) particulière(s) | 19 | Y a-t-il eu apparition d’effet(s) indésirable(s) ou une situation particulière depuis la dernière visite? | OuiNon |
| Arrêt définitif du traitement  | Arrêt définitif | 20 | Date de l'arrêt définitif de traitement  | jj/mm/aaaa |
| Raisons de l'arrêt de traitement  | 21 | Raisons de l'arrêt de traitement  | * Fin de traitement (définie dans le RCP)
* Survenue d’un effet indésirable suspecté d’être lié au traitement
* Progression de la maladie
* Effet thérapeutique non satisfaisant
* Décès
* Souhait du patient d'interrompre le traitement
* Patient perdu de vue
* Ne remplit plus les critères d’éligibilité
* Autre (préciser)
 |
| Si décès | 22 | Date de décès | jj/mm/aaaa |
| Si décès | 23 | Raison du décès | * Décès lié à un effet indésirable
* Décès lié à la progression de la maladie
* Autre raison (préciser)
 |

# Annexe 2 : Extrait du protocole d’étude

Objectifs de la recherche

Objectif principal

Comparer l’exactitude moyenne des données portant sur au moins 20 variables de deux méthodes de recueil de données :

* une extraction automatisée par IA à partir des DPI,
* et une saisie manuelle réalisée dans les conditions habituelles,

par rapport à la base de données de référence, chez des patients ayant initié un traitement dans l’un des AP pré-AMM inclus dans l’étude, après la fin de ces AP.

Objectif secondaire

Évaluer les performances de l'IA par rapport à une saisie manuelle dans la saisie de données recueillies dans le cadre d’accès précoces pré-AMM.

Comparer les performances des deux méthodes de recueil de données en termes :

1. d’exactitude des données par variable et par type de variable

2. d’erreur des données au global, par variable et par type de variable

3. de complétude des données au global, par variable et par type de variable

par rapport à la base de données de référence, pour les données d’au moins 20 variables

4. et évaluer la concordance des données au global, par variable et par type de variable entre les deux méthodes de recueil de données.

L’évaluation portera sur des patients ayant initié un traitement dans l’un des AP pré-AMM inclus dans l’étude, après la fin de ces AP.

Schéma de la recherche

Il s’agit d’une étude multicentrique, prospective et comparative.

Critères de jugement

Critère de jugement principal

L'objectif principal de l’étude sera évalué en comparant le taux moyen d’exactitude des données de l’étude par rapport à la base de données de référence de deux méthodes de recueil de données, une saisie manuelle et une extraction par IA.

Pour chaque dossier et méthode de recueil de données, le taux d’exactitude sera calculé de la façon suivante :

$$Taux d'exactitude=\frac{Nombre de données exactes du set prédéfini de variables}{Nombre de données du set prédéfini de variables}×100$$

Critères de jugement secondaires

Les critères de jugement secondaires permettant de comparer les performances d’une extraction de données par IA versus une saisie manuelle pour le recueil de données d’un set prédéfini de variables sont les suivants :

* Taux d’exactitude par variable et par type de variable (date, numérique, texte, etc.)
* Taux d’erreur parmi les données collectées au global, par variable et type de variable
* Taux de complétude des données attendues au global, par variable et type de variable
* Coefficient de concordance entre les données saisies manuellement et celles extraites par IA au global, par variable et type de variable.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. https://www.has-sante.fr/jcms/p\_3433891/fr/modele-de-protocole-d-utilisation-therapeutique-et-de-recueil-des-donnees-version-11-2023 [↑](#footnote-ref-2)