Protocole de recherche n'impliquant pas la personne humaine

TITRE	**Données d'Accès précoces Normalisées extraites par Traitement automatisé du langage en vue de leur Évaluation : étude comparative entre une saisie manuelle et une extraction via une intelligence artificielle**
TITRE COURT	**DANTE**
VERSION	**V1.0 du 10/04/2026**
PORTEUR DE PROJET	**Filière Intelligence Artificielle et Cancers (FIAC)**

Page de Signature du Protocole

Données d'Accès précoces Normalisées extraites par Traitement automatisé du langage en vue de leur Évaluation : étude comparative entre une saisie manuelle et une extraction via une intelligence artificielle

DANTE

PORTEUR DE PROJET **Filière Intelligence Artificielle et Cancers**	Nom : Marco FIORINI Fonction : Directeur Général Date : 13.04.2026 Signature Signed by: Marco Fiorini AF73BA9895C0421...
LABORATOIRES PHARMACEUTIQUES PARTICIPANT **AstraZeneca**	 Nom : Nicolas OZAN Fonction : Directeur Médical Oncologie Date : 13.04.2026 Signature Signé par : Nicolas Ozan 1099F18668EB4E7...
Amgen	Nom : Béatrice FIQUET Fonction : Directrice Médicale Onco-hématologie Date : 13.04.2026 Signature DocuSigned by: Béatrice FIQUET F16D06C3121049A...
MSD	Nom : Philippe BONNARD Fonction : Directeur Médical Oncologie Date : 13.04.2026 Signature Signé par : Philippe Bonnard C2307527478A4AF...

Sommaire

Liste des abréviations

AMM	Autorisation de mise sur le marché
ANSM	Agence nationale de sécurité du médicament
AP	Accès précoce
ARC	Attaché de recherche clinique
CRO	Société de recherche contractuelle / *Contract Research Organization*
DPI	Dossier patient informatisé
FIAC	Filière intelligence artificielle et cancers
HAS	Haute autorité de santé
HDH	Health Data Hub
IA	Intelligence artificielle
IC	Intervalle de confiance
Leem	Les entreprises du médicament
NLP	Traitement automatique du langage naturel / *Natural Language Processing*
PAS	Plan d'analyse statistique
PUT-RD	Protocole d'utilisation thérapeutique et de recueil de données
SIH	Système d'information hospitalier
TEC	Technicien d'études cliniques

Historique des mises à jour du protocole

Version et date	Sections concernées	Description et raison(s) de la mise à jour

Résumé de la recherche

Porteur de projet	Filière Intelligence Artificielle et Cancers (FIAC)
Laboratoires pharmaceutiques participant	AstraZeneca, Amgen, MSD
CRO	RCTs
TITRE LONG	Données d'Accès précoces Normalisées extraites par Traitement automatisé du langage en vue de leur Évaluation : étude comparative entre une saisie manuelle et une extraction via une intelligence artificielle
TITRE COURT	DANTE
JUSTIFICATION / CONTEXTE	Depuis la réforme de juillet 2021, la Haute Autorité de santé (HAS) encadre le dispositif d'accès précoce (AP) aux médicaments par un suivi structuré via des protocoles d'utilisation thérapeutique et de recueil de données (PUT-RD). Ces données en vie réelle, souvent les premières sur un traitement innovant dans l'indication de l'AP, sont essentielles pour évaluer efficacité, tolérance et conditions d'usage du médicament. Toutefois, malgré un dédommagement versé aux centres de santé pour le recueil de données, le taux d'exhaustivité des données reste limité (≈65 % contre un objectif de 90 %). Le recueil de données peut être perçu comme complexe et redondant par les professionnels de santé, en raison de la multiplicité des plateformes, de la densité des informations demandées et des ressources limitées. Ces contraintes entraînent hétérogénéité et variabilité de qualité entre centres. La HAS et l'organisation professionnelle des entreprises du médicament opérant en France (le LEEM) appellent à simplifier et optimiser le dispositif, notamment via l'usage de sources existantes et l'expérimentation de l'intelligence artificielle (IA). Les progrès récents en IA, en particulier le traitement automatique du langage naturel (NLP) et l'apprentissage automatique, permettent désormais d'extraire des données cliniques directement des dossiers patients informatisés (DPI). Cette approche pourrait réduire le fardeau des soignants, améliorer la qualité des données et en accélérer le recueil. L'étude DANTE propose de comparer, dans le cadre d'AP pré-autorisation de mise sur le marché (AMM), la qualité des données recueillies manuellement par les centres prescripteurs et celles extraites automatiquement par IA, en les confrontant à une base de référence, à savoir un recueil de données effectué par des attachés de recherche clinique (ARC) indépendants. L'évaluation portera sur les performances des deux méthodes de recueil de données par rapport à une saisie de référence en termes d'exactitude, d'erreur, et d'omission, ainsi que de concordance entre les deux méthodes, afin de déterminer si l'IA constitue une alternative fiable et efficiente au recueil manuel par les centres.
OBJECTIFS	**Objectif principal** Comparer l'exactitude 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 par les centres dans les conditions habituelles.

	Objectifs secondaires 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. d'omission des données au global, par variable et par type de variable 4. de concordance des données au global, par variable et par type de variable. L'évaluation portera sur des patients ayant initié le traitement en AP dans l'un des AP inclus dans l'étude, après la fin de ces AP.
SCHÉMA DE LA RECHERCHE	Étude multicentrique, prospective et comparative.
CRITÈRES D'INCLUSION	- Patient âgé de 18 ans ou plus - Patient dont la demande d'accès au traitement d'un AP participant à l'étude a été validée, dans l'un des centres participants.
CRITÈRES DE NON INCLUSION	- Patient s'étant opposé à la réutilisation à des fins de recherche de ses données personnelles collectées via la saisie manuelle dans les conditions habituelles - Patient s'étant opposé à l'utilisation ou à l'extraction par une IA de ses données personnelles figurant dans le DPI.
PROCÉDURES DE LA RECHERCHE	Deux méthodes de recueil sont évaluées : la saisie manuelle habituelle effectuée par les médecins prescripteurs ou leurs équipes et l'extraction automatisée par IA à partir des DPI. Elles seront comparées à des données de référence. Ainsi, trois bases de données sont constituées pour les mêmes patients : - Base manuelle : extraite des données saisies par les professionnels de santé dans le cadre des PUT-RD des AP participant à DANTE. Les données seront gérées par l'entreprise de recherche sous contrat (CRO) mandatée par le laboratoire pharmaceutique pour la mise en œuvre de l'AP (CRO primaire). Aucune intervention ne sera effectuée dans le cadre de DANTE. Les données seront collectées selon les pratiques habituelles des centres prescripteurs et leur qualité sera vérifiée selon les procédures habituelles que chaque laboratoire pharmaceutique met en œuvre pour ses AP. - Base IA : alimentée par l'extraction automatique des DPI. Les données seront gérées par chaque partenaire technologique chargé d'extraire les données des DPI des patients inclus dans l'étude. - Base de référence : saisie en aveugle par des ARC indépendants à partir des DPI. Les données seront gérées par une CRO contractée par la FIAC pour l'étude (CRO centrale). Les patients inclus auront initié un traitement prescrit dans le cadre de l'un des AP participant à l'étude. La participation est ouverte à tous les centres prescripteurs disposant d'une solution d'IA permettant d'extraire les données du set prédéfini de variables à partir des DPI. Les données analysées porteront sur un set prédéfini d'environ 20–25 variables. La CRO centrale assurera la fusion, le chaînage, la vérification et l'analyse statistique des données. L'étude respectera le Règlement Général sur la Protection des Données (RGPD) et des recommandations de la Commission nationale de l'informatique et des libertés (CNIL), avec pseudonymisation et sécurisation des transferts.
CRITÈRES DE JUGEMENT	**Critère principal :** Taux moyen d'exactitude Le taux moyen d'exactitude sera défini pour les deux méthodes de recueil par rapport à la base de données de référence, chez des patients ayant initié le traitement en AP dans l'un des AP pré-AMM inclus dans l'étude, après la fin de ces AP.

	Critères de jugement secondaires - Taux d'exactitude par variable et taux moyen d'exactitude par type de variable (date, numérique, catégorielle) - Taux moyen d'erreur et par type de variable, et taux d'erreur par variable - Taux moyen d'omission des données attendues au global et par type de variable, et taux d'omission par variable - Coefficient de concordance entre les données saisies manuellement et celles extraites par IA au global, par variable et type de variable.
TAILLE D'ÉTUDE	147 patients (tous AP confondus).
NOMBRE PRÉVU DE CENTRES	Pas de sélection des centres a priori. Tous les centres prescripteurs d'un traitement dans le cadre de l'un des AP participant à l'étude et disposant d'une solution d'IA 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 de recrutement (147 patients).
DURÉE PRÉVISIONNELLE	Durée prévisionnelle de la période d'inclusion : à déterminer. Durée prévisionnelle totale de la recherche : à déterminer.
ANALYSE STATISTIQUE DES DONNÉES	Population : ensemble des patients inclus dans l'étude. Sous-groupes d'analyse : Type de centres (centre hospitalier, centre hospitalier universitaire, centre de lutte contre le cancer, etc.). Analyse des critères principaux : Le taux moyen d'exactitude sera estimé avec son intervalle de confiance (IC) à 95 % pour l'extraction par IA et pour la saisie manuelle. Ces taux seront comparés à l'aide d'un test de Student pour données appariées ou, en cas de distribution non normale, d'un test de Wilcoxon pour données appariées. Analyse des critères secondaires L'ensemble des paramètres seront estimés avec leur IC à 95 % pour l'extraction IA et pour la saisie. Les taux seront comparés à l'aide d'un test de Mc Nemar et les taux moyen à l'aide d'un test de Student pour données appariées ou, en cas de distribution non normale, d'un test de Wilcoxon pour données appariées. La concordance entre les deux méthodes de recueil de données sera étudiée en deux niveaux : d'une part, via le kappa de Cohen pour chaque variable, afin de mesurer la concordance par variable, et d'autre part, via un kappa de Conger pour obtenir une estimation globale de la concordance, à la fois par type de variable et sur l'ensemble des dossiers/patients.

1 Introduction

1.1 Contexte de la recherche

Depuis la réforme de l'accès dérogatoire aux médicaments le 1er juillet 2021[1], la Haute Autorité de santé (HAS) est en charge des décisions d'autorisation d'accès précoce (AP), en collaboration étroite avec l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) pour les indications sans autorisation de mise sur le marché (AMM).

Lors de la demande d'autorisation d'AP, les laboratoires pharmaceutiques doivent soumettre un protocole établi selon un modèle spécifique de protocole d'utilisation thérapeutique et de recueil de données (PUT-RD)[2], qui sera revu par les autorités de santé, et au besoin modifié, avant d'être validé. Ce protocole est composé de plusieurs fiches qui doivent être renseignées par les médecins prescripteurs selon un calendrier de visites spécifique à chaque AP et correspondant à différents moments du suivi des patients dans l'AP : demande d'accès au traitement, instauration, suivi, et arrêt (éventuel) du traitement. L'étendue des données demandées varie selon les besoins des autorités de santé. Ainsi, un PUT-RD est davantage détaillé, notamment en matière de suivi du traitement, dans le cadre d'un AP pré-AMM que post-AMM.

L'analyse des données recueillies via les PUT-RD vise à permettre l'évaluation en continu des critères d'octroi de l'autorisation d'accès dérogatoire par les autorités de santé, et de contribuer à l'évaluation du médicament par la commission de la transparence en vue de leur inscription sur la liste des médicaments remboursables le cas échéant. Ces données sont souvent les premières concernant l'utilisation en vie réelle d'un traitement conformément à son indication validée par les autorités pour l'AP, notamment en matière d'efficacité et de tolérance, et sont à ce titre uniques et précieuses.

Les médecins et les pharmaciens souhaitant prescrire ou dispenser un médicament bénéficiant d'une autorisation d'accès dérogatoire sont tenus de participer au recueil des données et de les transmettre au laboratoire pharmaceutique. Les modalités d'organisation du recueil des données sont propres à chaque établissement, notamment selon les ressources humaines disponibles. Le recueil des données peut être délégué par le médecin prescripteur à un membre du personnel soignant, ou à un personnel de recherche clinique (technicien d'études cliniques (TEC) ou un attaché de recherche clinique (ARC)) s'il en dispose ce type de poste n'existant pas dans le soin. La responsabilité du recueil reste sous la responsabilité du médecin et du pharmacien hospitalier. Afin d'indemniser les établissements de santé pour le temps consacré par leurs équipes au recueil de données, une convention de dédommagement, dont le modèle a été publié par arrêté le 15 avril 2022[3], doit être établie entre le laboratoire pharmaceutique et chaque établissement de santé. Le dédommagement varie selon le taux d'exhaustivité des données calculé pour chaque établissement de santé. Avec ces mesures incitatives, l'objectif des autorités de santé est d'atteindre un taux d'exhaustivité des données de 90 %.

1.2 Justification de la recherche

Après quatre années de mise en œuvre, les retombées concernant le dispositif réformé d'AP aux médicaments sont positives en matière d'accès à la thérapeutique : 287 premières demandes d'autorisation d'AP ont été déposées par les industriels, dont 116 en oncologie-cancérologie, et 152

[1] Article 78 de la loi n°2020-1576 du 14 décembre 2020 de financement de la Sécurité sociale pour 2021.
[2] Modèle de protocole d'utilisation thérapeutique et de recueil des données (version novembre 2023), HAS
[3] Arrêté du 15 avril 2022 relatif au modèle de convention prévu aux articles R. 5121-70, R. 5121-74-5 et R. 5121-76-6 du code de la santé publique, Journal officiel

autorisations ont été délivrées[4]. Toutefois, les objectifs fixés concernant le recueil de données n'ont pas été atteints : selon l'organisation professionnelle des entreprises du médicament opérant en France (le Leem), le taux d'exhaustivité est estimé à 65 %[5], soit 25 points de moins que l'objectif fixé par la HAS. À ce jour, il n'existe pas de données concernant la qualité des données recueillies.

La collecte de données dans le cadre des PUT-RD pose plusieurs défis. Contrairement aux études cliniques qui reposent sur un engagement volontaire des professionnels de santé et dont la mission de recherche est explicite, les AP sont avant tout des dispositifs réglementaires dérogatoires encadrant la prescription, la délivrance et la prise en charge d'un médicament dans le cadre des soins courants. Bien que le recueil de données fasse partie des engagements des médecins prescripteurs et pharmaciens souhaitant prescrire ou dispenser un médicament en AP[6], la finalité du recueil de données dans ce contexte n'est pas toujours claire ou prioritaire pour les équipes soignantes. Par ailleurs, le recueil de données en supplément des informations enregistrées en routine dans le dossier des patients pour leur suivi médical peut apparaitre comme redondant pour les professionnels de santé, d'autant plus lorsque les ressources humaines sont limitées. La densité des données demandées dans les PUT-RD, en particulier pour les AP pré-AMM, et les exigences spécifiques à chaque AP, alourdissent les démarches pour les médecins prescripteurs. S'il reste possible de remplir les fiches au format papier, le plus souvent une plateforme numérique de saisie des données est spécifiquement développée par les laboratoires pour chaque PUT-RD. Il en découle une diversité de plateformes, chacune avec ses spécificités techniques et ergonomiques, qui complexifie la prise en main et l'accès aux outils de saisie par les personnes intervenant dans la saisie des données, et ce malgré la création de l'application Pasrel par l'Agence technique de l'information sur l'hospitalisation (ATIH) que tous les gestionnaires de plateformes doivent intégrer à leurs solutions[7]. L'ensemble de ces défis constitue un fardeau en termes de charge de travail et de ressources humaines qui ne favorise pas l'adhésion des médecins prescripteurs au recueil de données d'AP et est de nature à impacter leur qualité. Par ailleurs, l'hétérogénéité du fardeau du recueil et des habitudes de remplissage entre les centres en France, liées à la disponibilité du personnel pouvant effectuer la saisie, et variant selon le nombre de patients traités dans le cadre d'un AP, accentue la variabilité de la qualité des données.

Dans leur bilan d'octobre 2023, l'ANSM et la HAS[8] soulèvent la nécessité d'optimiser le recueil de données en proposant plusieurs axes d'amélioration, notamment en simplifiant le modèle de PUT-RD, et en recommandant l'utilisation de sources de données existantes (registres) pour limiter la redondance de saisie. En juin 2024, le Leem a publié 12 propositions pour optimiser et donner du sens au recueil de données[9], dont l'expérimentation du recours à l'intelligence artificielle (IA) pour collecter les données d'AP à partir de données existantes.

Les progrès récents en IA offrent des opportunités prometteuses pour améliorer les processus de collecte et de gestion des données tant dans les études cliniques qu'en vie réelle(Weissler et al. 2021; Quennelle et al. 2023). Si l'évaluation du recours à l'IA en recherche clinique est bien documentée et montre des capacités prometteuses(Han et al. 2024), son intégration en vie réelle pose plusieurs défis

[4] Haute Autorité de Santé - Comprendre l'évaluation des médicaments (consulté le 28/07/2025).
[5] Accès précoce : 12 propositions pour optimiser et donner du sens au recueil de données, Leem (juin 2024).
[6] Voir modèle de PUT-RD p.4.
[7] Les professionnels de santé disposent d'un compte Pasrel (Plateforme d'Accès aux SeRvices En Ligne) qui permet de se connecter via un identifiant unique à l'ensemble des outils numériques dédiés aux médicaments, y compris ceux mis à disposition par les laboratoires pour les accès dérogatoires.
[8] Accès précoce des médicaments : un bilan positif après deux ans de mise en place du dispositif, HAS (octobre 2023).
[9] Accès précoce : 12 propositions pour optimiser et donner du sens au recueil de données, Leem (juin 2024).

liés à l'hétérogénéité des pratiques en vie réelle(El Arab et al. 2025). Les outils d'IA, tels que ceux reposant sur le traitement automatique du langage naturel (NLP) et l'apprentissage automatique, sont de plus en plus utilisés pour automatiser et optimiser l'extraction d'informations médicales à partir des données générées en routine lors de la prise en charge des patients, organisées en dossiers patients informatisés (DPI) et stockées dans les systèmes d'information hospitaliers (SIH)(Hossain et al. 2023). Ces outils ont le potentiel d'accélérer le recueil des données, de réduire les erreurs humaines et d'améliorer la qualité des informations collectées, tout en réduisant le fardeau pour les professionnels et les établissements de santé(Weissler et al. 2021; Hom et al. 2022; Goodman et al. 2025; Aldea et al. 2025).

L'étude DANTE propose d'évaluer la qualité des données extraites des DPI de façon automatisée via une IA et celles saisies de façon manuelle par les médecins prescripteurs dans les conditions habituelles de leur pratique dans le cadre d'AP ayant initialement obtenu une autorisation au stade pré-AMM. Les données des deux types de saisie (IA et manuelles) seront vérifiées par rapport à une base de données de référence (*gold standard*) qui aura été saisie par des ARC indépendants à partir des DPI (données source). Les trois bases de données ainsi constituées (IA, manuelle, référence) concerneront les mêmes patients qui seront ensuite appariés pour permettre la comparaison des données. Il s'agira d'évaluer la qualité des données collectées manuellement et par une IA en termes d'exactitude, d'erreur, et d'omission par rapport aux données source, ainsi que de concordance entre les deux méthodes de recueil, afin d'évaluer si l'extraction par IA apporte une amélioration comparativement à une saisie manuelle, et si oui dans quelle mesure et de quelle nature.

L'étude DANTE portera sur des AP ayant obtenu une autorisation en phase pré-AMM, le volume de données demandées dans les PUT-RD étant plus importante dans ce type d'AP. Lorsque qu'un médicament en AP obtient son AMM dans l'indication de l'AP, le dispositif prévoit un passage possible d'AP pré-AMM à post-AMM avec un allègement ou non du PUT-RD sur décision de la HAS. Sous réserve que le PUT-RD demeure identique, impliquant ainsi que la collecte de données du set de données minimum perdure dans la phase post-AMM, l'étude pourra se poursuivre pour les AP concernés.

Les bases de données IA et de référence seront constituées à partir des informations disponibles dans les DPI jusqu'à la date de fin de collecte des données dans la base manuelle. Ainsi, les données enregistrées ultérieurement dans les DPI ne seront pas prises en compte, afin de garantir une comparaison équitable avec les deux modalités de recueil de données.

1.3 Retombées attendues

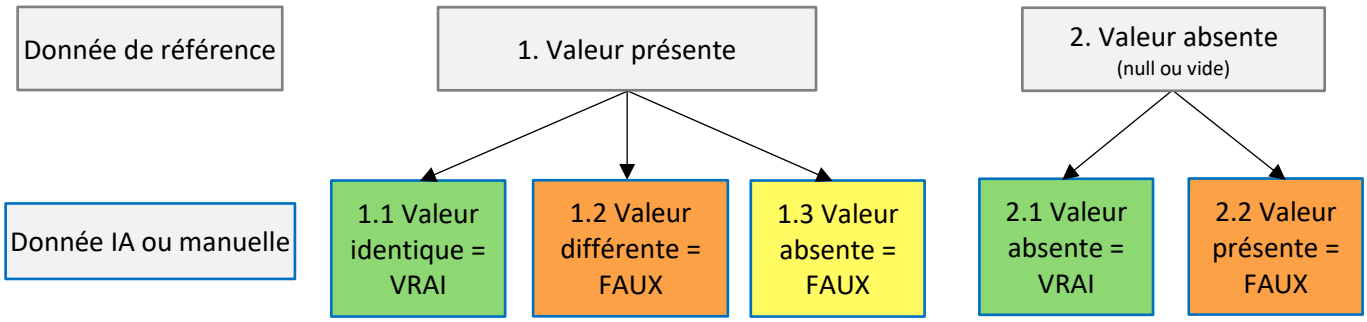
En documentant les premières utilisations de traitements innovants en dehors des essais cliniques, les données recueillies dans le cadre des AP fournissent des données précieuses sur l'usage en vie réelle de ces traitements (dans l'indication de l'AP), et sont d'une importance réglementaire, stratégique et scientifique majeure. Elles sont essentielles pour répondre aux exigences d'évaluation par les autorités de santé, cruciales pour les laboratoires responsables de la mise à disposition de ces traitements, et indispensables aux médecins prescripteurs pour comprendre le comportement du médicament chez leurs patients et mieux adapter leur prise en charge. En analysant et en comparant la saisie des données manuelle versus une saisie automatisée, cette recherche vise à créer des leviers pour améliorer la collecte et la qualité des données relatives aux médicaments innovants, afin que résultats de l'analyse de ces données puisse bénéficier à l'ensemble de l'écosystème.

Ayant pour porteur de projet la Filière Intelligence Artificielle et Cancers (FIAC), cette étude explore la dimension Filière et le CSF (Cloud, Security, et Filière) en analysant un cas d'usage spécifique de l'IA, particulièrement dans le domaine de l'oncologie. Cela pourrait permettre de mieux comprendre

comment intégrer efficacement l'IA dans l'extraction de données collectées dans le cadre du soin et enregistrées dans les DPI pour une utilisation en recherche ou surveillance des médicaments, tout en évaluant la fiabilité des informations obtenues et en définissant les tests de qualité nécessaires.

2 Objectifs de la recherche

Les indicateurs de performance utilisés dans cette étude ont été définis selon le schéma de classification des correspondances entre une donnée issue d'une saisie manuelle ou extraite par IA et une donnée de référence issue du DPI, présenté en Figure 1. Ce schéma précise les différentes situations de concordance ou de discordance par rapport aux données source, ainsi que les définitions associées des notions d'exactitude, d'erreur et d'omission qui serviront à l'évaluation des critères de jugement.



Définitions

Exactitude = 1.1 + 2.1	Correspondance exacte entre la valeur saisie manuellement ou extraite par IA et la valeur de référence, pour les valeurs de référence présentes ou absentes (null ou vides) **Taux d'exactitude = ((1.1 + 2.1) / (1+2)) * 100**
Erreur = 1.2 + 2.2	Différence entre la valeur saisie manuellement ou extraite par IA et la valeur de référence, pour les valeurs de référence présentes ou absentes (null ou vides) **Taux d'erreur = ((1.2 + 2.2) / (1+2)) * 100**
Omission = 1.3	Absence de valeur manuelle ou IA alors qu'une valeur de référence est présente **Taux d'omission = (1.3 / 1)*100**

Figure 1. Schéma de classification des correspondances entre une donnée manuelle ou IA et une donnée de référence, et définitions associées d'exactitude, d'erreur ou d'omission de donnée

En complément, la concordance des valeurs entre les deux modes de recueil des données est définie par la correspondance exacte entre la valeur manuelle et la valeur IA.

2.1 Objectif principal

Comparer l'exactitude 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 par les centres dans les conditions habituelles

chez des patients ayant initié un traitement dans l'un des AP inclus dans l'étude, après la fin de ces AP.

2.2 Objectifs secondaires

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

- d'exactitude des données par variable et par type de variable (date, numérique, catégorielle)
- d'erreur des données au global, par variable et par type de variable
- d'omission des données au global, par variable et par type de variable

et évaluer la concordance entre les données manuelles et les données IA au global, par variable et par type de variable.

L'évaluation portera sur des patients ayant initié le traitement en AP dans l'un des AP inclus dans l'étude, après la fin de ces AP.

3 Schéma de la recherche

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

3.1 Sélection des centres

L'étude est ouverte à tout établissement disposant d'une solution d'IA (en propre ou via un prestataire) permettant de collecter et structurer de façon automatisée les données d'un set prédéfini de variables à partir de DPI. Lorsqu'une demande d'accès à l'un des traitements des AP participant sera validée, le laboratoire pharmaceutique titulaire de l'autorisation d'AP transmettra les coordonnées du centre à la FIAC qui se chargera de vérifier que le centre dispose d'une solution d'extraction de données en propre ou via un prestataire Une convention sera établie avec les centres éligibles, avec la FIAC et/ou avec leur prestataire de solution d'IA le cas échéant. Autant de centres que nécessaire pour atteindre l'objectif de recrutement (147 patients) seront inclus dans l'étude. La FIAC travaillera en lien avec les développeurs de solution IA, en propre dans les centres ou prestataires externes des centres, désignés comme « partenaires technologiques » dans ce protocole.

3.2 Sélection des accès précoces

Les patients inclus seront issus d'au moins deux AP ayant initialement obtenu une autorisation au stade pré-AMM d'au moins deux des trois laboratoires pharmaceutiques ayant participé au développement de cette étude, AstraZeneca, MSD et Amgen. D'autres AP pré-AMM, initiés par ces trois laboratoires pharmaceutiques ou d'autres, pourront rejoindre l'étude au cours de sa mise en œuvre, pour atteindre le nombre de dossiers/patients nécessaires ou diversifier les sources de données.

3.3 Sélection des dossiers patient

Dans cette étude, l'unité d'analyse est le dossier du patient. Les critères d'inclusion et de non-inclusion ci-dessous concernent les patients dont les dossiers seront analysés.

3.3.1 Critères d'inclusion

- Patient âgé de 18 ans ou plus
- Patient dont la demande d'accès au traitement d'un AP participant à l'étude a été validée, dans l'un des centres participants.

3.3.2 Critères de non-inclusion

- Patient s'étant opposé à la réutilisation à des fins de recherche de ses données personnelles collectées via la saisie manuelle dans les conditions habituelles
- Patient s'étant opposé à l'utilisation ou à l'extraction par une IA de ses données personnelles figurant dans son DPI.

3.4 Critères de jugement

Les définitions des notions d'exactitude, d'erreur, d'omission et de concordance sont explicitées en Figure 1 (2.Objectifs de la recherche). Le set prédéfini de variables est décrit en Annexe 1.

3.4.1 Critère de jugement principal

Le critère principal d'évaluation permettant de comparer l'exactitude des données de l'étude entre une saisie manuelle et une extraction par IA sera la moyenne des taux d'exactitude de chaque dossier définit par :

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

Le taux moyen sera calculé pour les deux méthodes de recueil de données.

3.4.2 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 seront calculés pour les deux méthodes de recueil de données. Ils sont décrits de la façon suivante :

- Taux d'exactitude par variable défini par :

$$\frac{Nombre\ de\ dossiers\ dont\ la\ donnée\ de\ la\ variable\ \text{est}\ exacte}{Nombre\ de\ dossiers} \times 100$$

- Taux moyen d'exactitude par type de variable (date, numérique, texte, etc.), soit la moyenne des taux d'exactitude de chaque dossier défini par :

$$\frac{Nombre\ de\ données\ exactes\ d'un\ type\ de\ variables\ du\ set}{Nombre\ de\ \ variables\ d'un\ même\ type} \times 100$$

- Taux moyen d'erreur, soit la moyenne des taux d'erreur de chaque dossier défini par :

$$\text{Taux d'erreur} = \frac{Nombre\ de\ données\ erronées\ du\ set\ prédéfini\ de\ variables}{Nombre\ de\ variables\ du\ set\ prédéfini} \times 100$$

- Taux d'erreur par variable défini par :

$$\frac{Nombre\ de\ dossiers\ dont\ la\ donnée\ de\ la\ variable\ est\ erronée}{Nombre\ de\ dossiers} \times 100$$

- Taux moyen d'erreur par type de variable (date, numérique, texte, etc.), soit la moyenne des taux d'erreur de chaque dossier défini par :

$$\frac{Nombre\ de\ données\ erronées\ d'un\ type\ de\ variables\ du\ set}{Nombre\ de\ variables\ d'un\ même\ type} \times 100$$

- Taux moyen d'omission des données attendues, soit la moyenne des taux d'omission de chaque dossier défini par :

$$Taux\ d'omission = \frac{Nombre\ de\ données\ absentes\ du\ set\ prédéfini\ de\ variables}{Nombre\ de\ données\ attendues\ du\ set\ prédéfini\ de\ variables} \times 100$$

- Taux d'omission par variable défini par :

$$\frac{Nombre\ de\ dossiers\ dont\ la\ donnée\ de\ la\ variable\ est\ absente}{Nombre\ de\ dossiers\ dont\ la\ donnée\ de\ la\ variable\ est\ attendue} \times 100$$

- Taux moyen d'omission par type de variable : moyenne du taux d'omission par type de variable pour chaque dossier défini par :

$$\frac{Nombre\ de\ données\ \text{absentes}\ d'un\ type\ de\ variables}{Nombre\ de\ données\ attendues\ d'un\ même\ type\ de\ variable} \times 100$$

La concordance des données entre les deux modes de recueil sera évaluée de la façon suivante :

- Coefficient de concordance entre les données saisies manuellement et celles extraites par IA au global, par variable et type de variable.

3.5 Déroulé de la recherche

La recherche débutera au lancement du premier AP participant à DANTE, à savoir dès la validation de la première demande d'accès au traitement réalisée dans le cadre de cet AP. La période d'inclusion s'achèvera à l'inclusion du 147e patient. L'analyse finale des données commencera lorsque l'ensemble des données des 3 bases auront été validées.

4 Data management

4.1 Nature des données recueillies

Dans le cadre de ce projet, les données analysées portent sur un set composé d'environ 20 à 25 variables (Annexe 1). Ce choix méthodologique vise à restreindre le périmètre des données à un ensemble défini et limité, garantissant ainsi une approche objective et reproductible pour évaluer les performances de chaque mode de collecte de données. L'étendue des données demandées dans le cadre d'un PUT-RD d'un AP pré-AMM étant plus vaste que le set prédéfini pour cette étude, cette focalisation sur un set limité de variables permet de simplifier et standardiser l'évaluation pour différents AP.

Le set prédéfini de variables comprend des items issus des fiches de demande d'accès, d'instauration, de suivi (plusieurs visites possibles) et d'arrêt définitif de traitement figurant dans le modèle de PUT-RD. La sélection des variables a été réalisée par la FIAC et les laboratoires pharmaceutiques participants afin de déterminer les variables d'intérêt impératives, et garantir par leur diversité, en termes de complexité d'extraction, la robustesse du design de l'étude DANTE. L'étude permettra d'apporter des éléments en vue de la validation de l'utilisation d'une technologie de recueil de données par une IA dans le cadre des AP.

Les données collectées via les deux méthodes évaluées, manuelle et IA, et via des ARC pour la base de référence, concerneront les mêmes patients, et constitueront trois bases de données distinctes. Les données minimales nécessaires à l'appariement des patients seront collectées conformément aux spécifications définies pour assurer le chainage des patients entre les bases de données manuelle et IA afin de permettre la comparaison des deux méthodes de recueil de données.

4.2 Sources des données

L'étude repose sur une comparaison des performances de deux modes de recueil de données, manuel et IA, établies par rapport à une base de référence (pour l'exactitude, l'erreur, l'omission) ou entre elles (pour la concordance). Ainsi, trois bases de données seront utilisées :

1. Base de données manuelles : une extraction à partir de la base globale des données collectées prospectivement par les professionnels de santé dans le cadre des AP participant à cette étude sera réalisée par la CRO en charge de gérer l'AP, dite primaire dans ce protocole, sous la responsabilité du laboratoire pharmaceutique titulaire de l'autorisation d'AP. L'extraction comprendra uniquement les données correspondant au set prédéfini de variables (Annexe 1), et pour les patients inclus la base de données IA. Une extraction sera réalisée après chaque gel de base effectué en vue de la préparation des rapports de synthèse (intermédiaire ou final) qui doivent être remis aux autorités de santé selon une périodicité définie et à la fin de l'AP.
2. Base de données IA : cette base sera alimentée prospectivement par les données correspondant au set prédéfini de variables (Annexe 1), extraites par une solution IA à partir des DPI des patients dont la demande d'accès au traitement en AP a été validée par le laboratoire pharmaceutique titulaire de l'autorisation d'AP dans les centres sélectionnés. Le laboratoire devra fournir les identifiants des patients concernés pour permettre à l'IA de retrouver le patient dans les DPI. L'extraction des données portera sur les mêmes périodes que celles des données manuelles pour chaque AP participant. Cette base sera composée des données transmises par les partenaires technologiques, c'est-à-dire les développeurs de solution IA, en propre dans les centres ou prestataires externes des centres.
3. Base de données de référence : cette base sera constituée rétrospectivement par une saisie manuelle des données correspondant au set prédéfini de variables pour chaque patient notifié par le laboratoire pharmaceutique pour inclusion dans la base IA, indépendamment du succès d'inclusion ou non dans cette base. Seules les données disponibles dans les DPI à la même date que celle de fin de collecte manuelle et IA seront prises en compte, afin de garantir une comparaison équitable avec les deux modalités de recueil de données.

4.3 Méthode de recueil des données manuelles

Chaque laboratoire pharmaceutique participant à l'étude DANTE restera responsable de la collecte des données manuelles conformément au PUT-RD du médicament bénéficiant d'une autorisation d'AP. Cela comprend le choix de la CRO primaire, chargée de mettre en œuvre cette collecte de données selon les processus habituels de data management, notamment développement et mise à disposition d'une plateforme de recueil de données, suivi de la saisie des données et validation des données. Aucune intervention visant à modifier les pratiques de saisie de données d'AP par les médecins prescripteurs, ou de leurs équipes le cas échéant, n'est prévue dans le cadre de l'étude DANTE.

4.4 Méthode de recueil des données par IA

La collecte des données par IA a pour objectif :

- D'extraire automatiquement les données requises par le set prédéfini de variables depuis les DPI des patients inclus

- De comparer la qualité des données ainsi extraites à celle du recueil manuel habituel
- De valider la faisabilité, l'exactitude et la généralisabilité de cette méthode auprès de plusieurs établissements.

Cette approche repose sur le traitement de données pseudonymisées et sur la collaboration entre les partenaires technologiques développant les solutions IA et la CRO centrale supervisée par la FIAC. Les droits d'accès aux données, la pseudonymisation, l'archivage des sorties et les modalités de partage seront encadrés par les conventions tri ou quadripartites entre centres et/ou partenaires technologiques, la FIAC, et la CRO centrale. Afin de garantir une comparaison équitable entre les méthodes de recueil de données, les partenaires technologiques ne vérifieront l'exactitude des données extraites par leurs algorithmes dans les DPI, ni ne les modifieront. À cette fin, chaque partenaire technologique se sera engagé à documenter son algorithme et enregistrer les logs d'extraction pour traçabilité.

La saisie des données par IA suivra les étapes suivantes :

1. **Accès aux données et périmètre technique**

Les partenaires technologiques devront disposer d'un accès autorisé aux données nécessaires, via :

- Les bases structurées des DPI (prescriptions, bilans biologiques, admissions, etc.),
- Les textes non structurés (comptes rendus médicaux, lettres, notes d'évolution),
- Éventuellement les métadonnées des systèmes d'imagerie (PACS).

Chaque partenaire technologique validera en amont la possibilité d'accéder aux sources nécessaires dans chaque centre.

2. **Identification de la population cible**

L'algorithme devra commencer par identifier les patients ayant reçu le médicament concerné dans le cadre de l'AP, via :

- Le nom du traitement dans les prescriptions,
- Des marqueurs spécifiques dans les notes ou fiches d'instauration,
- Ou les identifiants pseudonymisés transmis par les laboratoires pharmaceutiques.

3. **Développement ou adaptation de l'algorithme IA**

L'extraction reposera sur des technologies de NLP et/ou d'apprentissage automatique. Deux cas de figure pourront se présenter selon le partenaire technologique :

- Approche supervisée : l'algorithme est entraîné sur un corpus annoté manuellement.
- Approche par règles ou hybride : dans les cas plus simples, une extraction peut être réalisée par patterns, expressions régulières ou modèles de langage.

L'algorithme sera conçu pour extraire seulement les variables du set prédéfini (Annexe 1).

Les données utilisées pour le développement et l'amélioration de l'algorithme devront être distinctes des données de l'étude DANTE. En effet, il conviendra de ne pas utiliser les mêmes données pour l'entraînement de l'algorithme et pour l'évaluation des critères de jugement de l'étude afin de ne pas surestimer les performances réelles de l'algorithme.

4. **Traitement et structuration des données**

Les données extraites seront formatées de façon structurée, avec un mapping vers un modèle commun de données lorsque possible (OMOP-CDM préférentiellement). Cela permettra :

- Une homogénéité inter-centres

- Une traçabilité de chaque donnée extraite
- Une compatibilité avec les outils d'agrégation pour constituer une base commune en vue de l'analyse des données.

5. Transmission et comparaison

À la fin des AP participant, les données seront transmises à la CRO centrale via un canal sécurisé, au format standardisé qui sera défini dans un plan de data management. Elles seront appariées par dossier patient pour évaluer les critères de jugement principal et secondaires de l'étude.

4.5 Méthode de recueil de données de référence

Des ARC indépendants seront contractualisés par la FIAC, via la CRO centrale, pour effectuer la saisie des données de référence. Un masque de saisie correspondant au set prédéfini de variables sera réalisé par la CRO centrale. La saisie sera réalisée en aveugle, c'est-à-dire sans accès préalable aux données collectées par l'une ou l'autre des deux méthodes évaluées.

5 Aspects statistiques

5.1 Calcul de la taille d'échantillon

Un total de 128 patients/set de variables est nécessaire pour détecter une différence moyenne du taux d'exactitude d'au moins 5 % (si elle existe) entre les 2 types de saisies avec une variabilité de cette différence de 20 %. L'analyse est basée sur un test t pour données appariées (un même dossier étant utilisé pour les deux lectures), avec un risque alpha de 0.05 et une puissance de 80 %.

En considérant 15 % de patients à exclure des analyses (suivi sur plusieurs centres ou patient n'ayant pas initié le traitement), 147 (128+19) patients/set de variables sont nécessaires afin de pouvoir répondre à l'objectif principal.

Afin d'inclure un nombre suffisant d'opérateurs pour faire la saisie manuelle (ARC/TEC, médecin, etc.), la variabilité inter-saisie tend à se neutraliser lorsque ceux-ci sont répartis de manière équilibrée sur un large échantillon. De ce fait, l'impact des différences systémiques entre les opérateurs qui saisissent sur les résultats globaux est minimisé. Cette approche repose sur l'hypothèse que les biais inter-saisie sont hétérogènes et non corrélés, permettant ainsi de diluer leur influence à mesure que le nombre d'opérateurs qui saisissent augmente.

Pour éviter un effet centre dans les analyses, aucun centre ne devra inclure plus de 15 % de patients (seuil empirique). Cette limite vise à garantir une répartition équilibrée des données entre les centres et à minimiser le risque qu'un centre exerce une influence disproportionnée sur les résultats.

5.2 Méthodes statistiques employées

5.2.1 Considérations générales

L'exploitation statistique ne débutera qu'après vérification de la validité de la base de données globale, résultant de la fusion de la base de données manuelles, de la base de données IA, et de la base de données de référence. La base de données globale sera alors gelée. Après le gel de la base de données, les analyses statistiques seront conduites par un biostatisticien qualifié, à l'aide du logiciel SAS® (V9.4 ou ultérieures) ou du logiciel R (CRAN).

Un plan d'analyse statistique (PAS) détaillé sera défini et validé avant le gel de la base de données.

Considérations statistiques

- Les variables qualitatives seront décrites en utilisant le nombre de données renseignées et manquantes et, pour chaque modalité, la fréquence et le pourcentage (en référence aux données renseignées),
- Les variables quantitatives seront décrites en utilisant le nombre de données renseignées et manquantes, la moyenne, l'écart type, la médiane, les 1er et 3ème quartiles, le minimum et le maximum,
- Le risque de première espèce alpha sera fixé à 5 %.

Gestion des données manquantes

Aucune imputation des données manquantes n'est envisagée.

Multiplicité des tests

Aucun ajustement des risques d'erreur n'est prévu.

5.2.2 Population d'étude

- Dossiers des patients ayant reçu au moins une dose du traitement dans le cadre de l'AP

- Dans un des centres ayant une solution IA implantée (en propre ou via un prestataire).

Les analyses pourront être présentées selon le type de centres (centre hospitalier, centre hospitalier universitaire, centre de lutte contre le cancer, etc.), sous réserve d'un nombre de dossiers suffisant.

5.2.3 Critère de jugement principal

Le taux d'exactitude sera estimé avec son intervalle de confiance pour l'extraction IA et pour la saisie manuelle par rapport à la base de référence. Ces taux seront comparés à l'aide d'un test de Student pour données appariées ou, en cas de distribution non normale, d'un test de Wilcoxon pour données appariées.

5.2.4 Critères de jugements secondaires

L'ensemble des données seront préalablement décrites en termes de caractéristiques générales, incluant notamment la présence de valeurs manquantes, leur répartition et la qualité globale des réponses.

L'ensemble des paramètres seront estimés avec leur intervalles de confiance (IC) à 95 % pour l'extraction IA et pour la saisie. Les taux seront comparés à l'aide d'un test de Mc Nemar et les taux moyen à l'aide d'un test de Student pour données appariées ou, en cas de distribution non normale, d'un test de Wilcoxon pour données appariées.

La proportion de données concordantes sera estimée avec son IC à 95 %. La concordance entre les deux méthodes sera étudiée en deux niveaux. D'une part, via le kappa de Cohen pour chaque variable, afin de mesurer la concordance par variable. Et d'autre part, via un kappa de Conger pour obtenir une estimation globale de la concordance, à la fois par type de variable et sur l'ensemble des dossiers.

6 Déroulement de la recherche

6.1 Aspects réglementaires

L'étude DANTE sera menée dans le respect du Règlement Général sur la Protection des Données (RGPD) et des recommandations de la Commission nationale de l'informatique et des libertés (CNIL). Les traitements de données à caractère personnel effectués pour les besoins de l'étude seront

encadrés par le référentiel RS-003 relatif aux données collectées dans le cadre des AP et par la méthodologie de référence MR-004 relative aux recherches impliquant des données de santé.

Pour les données saisies manuellement dans le cadre habituel de la mise en œuvre de l'AP, les données directement identifiantes demeureront dans les centres prescripteurs. Les patients seront pseudonymisés selon le référentiel RS-003 avant tout transfert. Les CRO primaires, mandatées par les laboratoires pharmaceutiques, assureront la collecte et la pseudonymisation des données issues du recueil manuel.

Pour les données extraites par IA, chaque partenaire technologique sera responsable de la pseudonymisation et de la transmission des données issues des SIH. Chacun sera également responsable de l'information des patients inclus et de recueillir leur opposition le cas échéant.

La base de référence sera constituée conformément à la MR-004 par la CRO centrale. Par ailleurs, la CRO centrale recevra les bases manuelles et IA pseudonymisées. Le chaînage des patients nécessaire à la constitution de la base fusionnée pour analyse sera réalisé grâce au pseudonyme unique présent dans les trois bases et conforme au référentiel RS-003. Une analyse d'impact relative à la protection des données (AIPD) sera conduite et validée, afin d'identifier les risques et définir les mesures correctrices appropriées.

6.2 Rôles et responsabilités

Parmi les acteurs impliqués dans la mise en œuvre de cette étude, les deux premiers sont indépendants de l'étude et gérés par les laboratoires titulaires des autorisations d'accès précoces participant à l'étude, et les deux derniers sont gérés par la FIAC :

- Les CRO primaires responsables de la mise en œuvre de la collecte et de la gestion des données du PUT-RD auront pour missions :
    - D'organiser la collecte et la saisie des données dans le cadre du PUT-RD via une plateforme numérique, selon le cahier des charges défini avec le laboratoire et sans intervention spécifique en lien avec l'étude
    - D'assurer les activités de data management conformément aux exigences réglementaires et du cahier des charges défini avec le laboratoire concernant le contrôle qualité des données
    - D'extraire et pseudonymiser les données correspondant au set de données défini dans le cadre de ce protocole pour les patients inclus dans le bras IA
    - De transférer la base de données ainsi constituée au laboratoire.

- Les centres participant aux AP seront responsables de la saisie des données de chaque PUT-RD via les médecins prescripteurs, ou leurs équipes soignantes et de recherche le cas échéant, selon leurs procédures, ressources et modes de fonctionnement interne, sans intervention spécifique dans le cadre de l'étude. Le contrôle qualité des données saisies manuellement relève de la responsabilité et des procédures de chaque laboratoire.

- Les partenaires technologiques auront pour missions :
    - De développer et entrainer leurs algorithmes sur d'autres données que celles utilisées dans l'analyse
    - De documenter l'algorithme final et enregistrer les logs d'extraction pour traçabilité

  - D'extraire les données des patients correspondant au set prédéfini via l'IA sans vérification manuelle, ni ajout manuel de données
  - De pseudonymiser les données collectées
  - De transférer de façon sécurisée les données structurées à la CRO centrale.

- Une CRO centrale, tiers de confiance, aura pour missions :
  - De fusionner les bases de données manuelles transmises par les CRO primaires
  - De chainer les données saisies manuellement avec celles extraites par IA
  - D'assurer la gestion des données de la base fusionnée
  - De constituer une base de référence à partir des DPI par la saisie de données réalisée par des ARC dans les centres participants
  - De définir un PAS et réaliser les analyses.

6.3 Sécurité et confidentialité des données

La protection des données sera assurée par des mesures techniques et organisationnelles conformes aux standards en vigueur. Les transferts entre partenaires s'effectueront via des canaux sécurisés (TLS 1.2 ou supérieur, SFTP, HTTPS). Les CRO primaires et les partenaires technologiques demeureront responsables de la sécurité et de l'hébergement de leurs données. La CRO centrale sera garante de la sécurité des données qui lui seront transférées et de la base de référence, et devra en confier l'hébergement à un Hébergeur de Données de Santé (HDS) certifié et localisé en France.

Les accès aux données hébergées par la CRO centrale sera limité aux seules personnes habilitées, selon un principe de droits minimaux et d'authentification renforcée. L'ensemble des opérations sera tracé par un journal d'audit. Chaque partenaire technologique conservera les journaux d'extraction de son algorithme, garantissant la traçabilité complète du processus.

Ces dispositions garantissent la confidentialité, l'intégrité et la disponibilité des données tout au long de l'étude.

6.4 Contrôle qualité

Cette étude repose sur la collecte de données dans le cadre d'AP pour évaluer une méthode innovante de collecte de données. Elle compare les performances (exactitude, erreur, omission) de la collecte effectuée par une solution d'IA avec la saisie manuelle effectuée dans les conditions habituelles (sur une plateforme numérique dédiée à l'AP) par le médecin prescripteur ou un membre de son équipe soignante ou de recherche (TEC/ARC) le cas échéant, selon l'organisation propre à chaque établissement participant. Aucune intervention ne sera effectuée dans le cadre de l'étude auprès des centres participant aux AP et chargés de recueillir les données manuellement, afin de garantir le cadre d'une étude menée en vie réelle non interventionnelle et de ne pas biaiser les résultats. Chaque laboratoire ou CRO primaire pourra mettre en œuvre un contrôle qualité des données manuelles selon leurs procédures habituelles dans le cadre de la mise en œuvre de leur AP et sans lien avec l'étude DANTE. Le cas échéant, le processus et les actions menées seront documentés par les laboratoires et pris en compte dans le rapport final de l'étude.

La saisie des données par IA sera réalisée en aveugle, sans intervention humaine pour réviser les données structurées par rapport aux documents source, afin d'éviter tout biais qui pourrait améliorer artificiellement les indicateurs de performance. À cet effet, les algorithmes seront documentés par chaque partenaire technologique et les logs d'extraction des données permettront d'en garantir la traçabilité. Tout écart sera décrit et son impact évalué dans le rapport final de l'étude.

Tout au long du recueil de données, la FIAC vérifiera régulièrement les données extraites par IA avec chaque prestataire technologique afin de s'assurer de leur pertinence au regard du set prédéfini de variables. Au premier gel de base des données manuelles réalisé pour les besoins du premier rapport périodique de synthèse destiné à la HAS[10], une vérification de la concordance des données extraites par IA par l'ensemble des partenaires technologiques sera effectuée par la CRO centrale pour tous les dossiers des patients inclus à date dans l'étude afin de s'assurer de la qualité et de la cohérence des données extraites. Cette vérification permettra de s'assurer du bon déroulement du projet et d'envisager les ajustements nécessaires le cas échéant.

7 Limites et points forts de l'étude

Biais de couverture des données et de définition des variables

Les données demandées dans les PUT-RD sont formulées et organisées d'une manière qui leur est propre. Si la plupart des informations demandées font partie du suivi habituel des patients et à ce titre figurent dans les DPI, certaines peuvent toutefois ne pas y avoir de correspondance exacte (ex. le motif d'arrêt de traitement) et d'autres n'y figurent pas du tout (ex. critère d'efficacité auto-rapporté par le patient). C'est pour éviter cet écueil que les données analysées dans le cadre de DANTE porteront uniquement sur un set prédéfini de variables, et que ce set a été revu par des partenaires technologiques afin de valider la faisabilité du recueil de ces données à partir des DPI. L'étude DANTE constitue une première étape d'évaluation de l'intérêt et de la faisabilité de l'extraction par IA de données d'AP à partir des DPI, dont les enseignements seront précieux pour envisager une automatisation future du recueil de données dans le cadre des AP.

Biais lié à la qualité des données sources et à l'absence de vérification des données source dans le cadre des AP

Les données de la base manuelle sont saisies par les médecins prescripteurs ou leurs équipes dans les conditions habituelles et spécifiques aux AP qui ne sont pas celles de la recherche biomédicale. De ce fait, les laboratoires pharmaceutiques ne sont pas promoteurs d'une recherche dans le contexte des AP et ne peuvent mandater du personnel pour effectuer un contrôle qualité sur site. Les données d'AP sont saisies directement dans les formulaires du PUT-RD, au format papier ou le plus souvent sur une plateforme dédiée. La traçabilité des données n'est pas exigée dans ce contexte, tandis que les données extraites par IA ou de la base de référence proviennent de la même source, les DPI. L'étude DANTE n'évalue pas la véracité intrinsèque des DPI. Comme pour toute source clinique, les informations contenues dans les DPI peuvent être incomplètes ou comporter des erreurs (ex. mauvaise sélection dans un menu déroulant, saisie erronée ou dans un champ inapproprié). Ces erreurs seront donc strictement reproduites dans les données extraites par IA, et probablement également dans les données de référence, même si la lecture manuelle des DPI par des ARC expérimentés pourra en rectifier certaines (ex. vérification auprès de l'équipe médicale, recoupement d'une information présente en plusieurs endroits du DPI). Ainsi, on ne peut exclure le risque que des discordances entre les données manuelles et les données de référence soient à tort attribuées à une erreur de saisie manuelle. Afin de renforcer la validité de la base de référence, une attention particulière sera portée à ces discordances lors de la revue des données avant analyse avec un retour vers le centre si nécessaire pour déterminer la valeur de référence correcte. L'analyse de la complétude des données

[10] Les laboratoires pharmaceutiques titulaires de l'autorisation d'AP sont tenus de fournir aux autorités de santé un rapport de synthèse comprenant toutes les informations recueillies dans le cadre de la mise en œuvre du PUT-RD. Le premier rapport est généralement attendu neuf mois après l'autorisation d'AP, avec un gel de base pouvant survenir jusqu'à deux mois avant la date de remise du rapport.

devrait être moins concernée par ce type d'erreurs. Une donnée manquante dans la base manuelle ou IA et manquante dans la base de référence sera considérée comme concordante.

Biais de classification lié à la construction de la base de référence

La base de référence sera constituée par du personnel expérimenté et formé spécifiquement à l'étude mais les erreurs de lecture des DPI et de saisie ne pourront être exclues, et ce malgré des procédures de contrôle à la saisie (ex. détection de valeur en dehors des bornes définies). En l'absence de ressources suffisantes pour mettre en place une double saisie indépendante, la vérification auprès des centres des discordances identifiées lors de la comparaison des trois bases (IA/manuel, IA/référence et manuel/référence) constitue une mesure pertinente pour renforcer la fiabilité des données de référence.

Biais de représentativité et hétérogénéité inter-centres et inter-technologies

L'étude DANTE se limite à un nombre restreint de centres participants, ce qui peut affecter la représentativité des résultats. Les différences dans les systèmes de DPI utilisés et dans la structuration des données peuvent également engendrer des variations de qualité des données entre centres. De même, les algorithmes utilisés seront propres à chaque partenaire technologique. Par ailleurs, la variabilité des pratiques de saisie inter-médecins prescripteurs introduit une forte hétérogénéité de qualité selon les utilisateurs. Plusieurs précautions ont été prises pour limiter l'impact de ces variabilités technologiques ou humaines, comme le fait d'inclure des patients issus d'au moins deux AP, d'avoir sélectionné des variables de différents types, et de viser une diversité de centres. Ces variabilités seront documentées dans la mesure du possible, l'objectif de DANTE étant aussi de refléter la pratique réelle, diverse par nature.

Biais de sélection des centres lié aux capacités d'extraction automatisée par IA

Les centres participants à l'étude DANTE seront sélectionnés sur la base de leur capacité à extraire de manière automatisée les données cliniques à l'aide de technologies d'IA, en propre ou via un prestataire externe. En conséquence, certains centres participant à des AP inclus dans le périmètre de l'étude pourraient ne pas être éligibles à DANTE.

Ce choix méthodologique implique un biais de sélection des centres, potentiellement en faveur de structures disposant de capacités numériques et organisationnelles plus avancées, notamment des centres de plus grande taille. Toutefois, l'objectif principal de l'étude DANTE n'est pas d'être représentative de l'ensemble des centres impliqués dans les dispositifs d'AP, mais d'évaluer la faisabilité, la qualité et la robustesse de l'extraction automatisée de données par IA dans ce cadre spécifique. Lors de la mise en œuvre de l'étude, une diversité de centres sera recherchée mais la priorité sera donnée à l'atteinte de l'objectif de recrutement des patients, de façon à permettre l'évaluation opérationnelle de la technologie.

Les enseignements issus de l'étude DANTE ont vocation à être généralisables dès lors que des conditions comparables de structuration, de qualité et d'accessibilité des données en amont sont réunies. À ce titre, l'étude permettra également d'identifier et de documenter les conditions nécessaires à une mise en œuvre efficace de l'extraction automatisée de données par IA, conditions susceptibles d'être appliquées ultérieurement à l'ensemble des centres, indépendamment de leur taille, et à d'autres contextes cliniques présentant des caractéristiques similaires.

8 Publication des résultats et valorisation

Les résultats seront présentés sous forme de rapport d'étude, qui sera partagé aux laboratoires participant à DANTE et aux membres du comité scientifique. Les résultats de l'étude seront par ailleurs transmis au Health Data Hub (HDH) afin d'être intégrés dans le répertoire des projets du site du HDH.

Des communications externes seront réalisées afin de valoriser les résultats de l'étude. Cela pourra consister et sans se restreindre, en un manuscrit scientifique, ainsi que des abstracts / posters qui seront proposés pour présenter les résultats. Le choix de la revue et des potentiels congrès ciblés sera défini avec le comité scientifique.

9 Références bibliographiques

Aldea, M., L. Zullo, V. Levrat, et al. 2025. « Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions ». *Annals of Oncology* 0 (0). https://doi.org/10.1016/j.annonc.2025.12.006.

El Arab, Rabie Adel, Mohammad S. Abu-Mahfouz, Fuad H. Abuadas, et al. 2025. « Bridging the Gap: From AI Success in Clinical Trials to Real-World Healthcare Implementation—A Narrative Review ». *Healthcare* 13 (7): 701. https://doi.org/10.3390/healthcare13070701.

Goodman, Keith, Chris Cook, Dani Weatherbee, et al. 2025. « Automating Data Entry from Electronic Health Record to Electronic Data Capture Using Trusted Cloud-Based Application in Multisite Cancer Clinical Trials ». *Journal of the Society for Clinical Data Management* 5 (1). https://doi.org/10.47912/jscdm.371.

Han, Ryan, Julián N. Acosta, Zahra Shakeri, John P. A. Ioannidis, Eric J. Topol, et Pranav Rajpurkar. 2024. « Randomised Controlled Trials Evaluating Artificial Intelligence in Clinical Practice: A Scoping Review ». *The Lancet Digital Health* 6 (5): e367-73. https://doi.org/10.1016/S2589-7500(24)00047-5.

Hom, Julie, Janet Nikowitz, Rebecca Ottesen, et Joyce C. Niland. 2022. « Facilitating Clinical Research through Automation: Combining Optical Character Recognition with Natural Language Processing ». *Clinical Trials (London, England)* 19 (5): 504-11. https://doi.org/10.1177/17407745221093621.

Hossain, Elias, Rajib Rana, Niall Higgins, et al. 2023. « Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review ». *Computers in Biology and Medicine* 155 (mars): 106649. https://doi.org/10.1016/j.compbiomed.2023.106649.

Quennelle, Sophie, Maxime Douillet, Lisa Friedlander, et al. 2023. « The Smart Data Extractor, a Clinician Friendly Solution to Accelerate and Improve the Data Collection During Clinical Trials ». *Studies in Health Technology and Informatics* 302 (mai): 247-51. https://doi.org/10.3233/SHTI230112.

Weissler, E. Hope, Tristan Naumann, Tomas Andersson, et al. 2021. « The role of machine learning in clinical research: transforming the future of evidence generation ». *Trials* 22 (1): 537. https://doi.org/10.1186/s13063-021-05489-x.

Annexes

Annexe 1. Liste des variables issues du modèle de PUT-RD de la HAS et comparées dans le cadre du projet DANTE

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?	Oui Non
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

Fiche du PUT-RD	Catégorie de la fiche du PUT-RD	N° et description de la variable	Valeurs attendues
			- 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)