

## ***L'IA associée à l'expertise humaine va transformer la collecte de données en recherche clinique, selon une étude publiée dans Annals of Oncology***

*L'IA performe dans l'extraction et la structuration de données issues de comptes rendus médicaux et surperforme quand elle est associée à l'expertise humaine.*

**Paris, le 18 décembre 2025** – Gustave Roussy, premier centre de lutte contre le cancer en France et en Europe, et Lifen, leader français de l'intelligence des données de santé, annoncent les résultats d'une étude multicentrique sur l'automatisation de l'extraction de données de comptes rendus médicaux, publiés dans *Annals of Oncology*. Cette étude démontre que l'IA permet, à partir de données très hétérogènes, d'accélérer la constitution d'une grande base de données de manière homogène, fiable et sécurisée, accélérant ainsi la compréhension des maladies et le développement de nouveaux traitements. Ces résultats ouvrent la voie à un changement de paradigme dans la recherche clinique. L'étude s'inscrit dans le cadre de la cohorte Lucc (*Large & Unified Cancer Cohort*), une initiative française soutenue par France 2030 et portée par Lifen et Gustave Roussy.

Dans le cadre de la recherche clinique, les bases de données sont aujourd'hui majoritairement renseignées par des experts de la data que sont les techniciens d'études cliniques (TEC) et attachés de recherche clinique (ARC). Ces informations proviennent principalement des dossiers patients, des comptes rendus médicaux et des examens biologiques ou d'imagerie. Cependant, l'extraction des données n'est pas simple car la majorité des dossiers de santé sont stockés sous des formats non structurés, rédigés en langage naturel, avec une terminologie hétérogène, des abréviations, des styles et des niveaux de qualité qui diffèrent d'un centre à l'autre. Une vérification de la saisie des données sensibles est systématiquement effectuée par un ARC superviseur dans des essais cliniques prospectifs.

Le projet Lucc vise à constituer des bases de données cliniques à grande échelle en automatisant l'extraction des données médicales via l'intelligence artificielle.

L'étude rétrospective a été menée sur 311 patients et 31 variables cliniques (démographiques, facteurs de risque, biomarqueurs génomiques, traitements) issues de 10 centres participants (établissements de santé publics et privés). Elle a comparé trois méthodes d'extraction de données depuis des comptes-rendus médicaux : la méthode manuelle où des attachés de recherche clinique expérimentés ont extrait manuellement les données via une plateforme électronique ; la méthode automatisée par IA utilisant l'intelligence artificielle pour une extraction automatique ; et la méthode hybride combinant extraction par IA, ciblage par IA et révision manuelle complémentaire.

Les résultats démontrent que :

- L'IA seule surpasse systématiquement l'approche manuelle sur chacune des 31 variables cliniques étudiées et pour chacun des 10 centres participants.
- L'IA seule réduit notamment de moitié les erreurs par rapport à la saisie manuelle (taux d'erreur de 7,0% pour l'IA contre 14,2% pour la méthode manuelle). Étant dans le cadre d'une étude rétrospective, la vérification humaine par un ARC superviseur n'a pas été faite.
- L'IA réduit la variabilité entre centres participants par rapport à la structuration manuelle. Cela signifie qu'elle est capable de mieux systématiser la façon dont les données sont collectées et permet de répondre à l'enjeu d'homogénéité des données dans les études multicentriques.
- Et surtout, la méthode hybride IA/expertise humaine va encore plus loin : elle combine l'extraction par l'IA avec une révision manuelle ciblée par l'IA sur les 30 % des cas qu'elle juge les plus incertains, ce qui fait chuter le taux d'erreur à 4,4 %, tout en conservant une vitesse de traitement quatre fois supérieure à l'analyse strictement manuelle.

Les résultats de cette étude sont inédits. Ils laissent entrevoir que l'IA peut redéfinir la manière de travailler des équipes de recherche clinique, qui vont pouvoir se concentrer sur des tâches à plus haute valeur ajoutée.

En automatisant l'extraction de données médicales, l'IA réduit le taux d'erreur, accélère le traitement des données et garantit une précision et une homogénéité essentielles pour gérer de grands volumes de données. L'approche par IA augmentée par l'expertise humaine est plus rapide mais aussi plus performante et efficiente. Elle permet également d'inclure des centres plus petits, souvent exclus faute de ressources, enrichissant ainsi la diversité des patients dans les études multicentriques. Ces avancées facilitent la conduite de projets de recherche de grande ampleur et ouvrent des perspectives prometteuses pour la recherche clinique française.

### **Méthodologie**

- Parmi les 10 000 patients de la cohorte LUCC Cancer du poumon, 311 patients issus de 10 centres ont été tirés au sort.
- Critères d'inclusion pour chaque patient : avoir au moins 5 comptes-rendus médicaux en français (hors résultats de laboratoire, imagerie et convocations) ; absence d'antécédent oncologique ou de cancer concomitant.
- Trois approches ont été évaluées : extraction manuelle par des professionnels de la recherche clinique, extraction automatisée par l'IA, et approche hybride combinant IA et révision humaine.
- Chaque groupe a travaillé sur les mêmes documents pseudonymisés, en suivant des règles d'extraction identiques et en appliquant des contrôles de cohérence automatiques.
- Pour chaque patient et chaque variable, les valeurs extraites par les différentes méthodes ont été comparées.
- Les valeurs concordantes entre méthodes ont été considérées comme justes.
- En cas de divergence, un professionnel senior de recherche clinique, supervisé par des oncologues et aveugle à la méthode utilisée, attribuait la valeur correcte. La valeur finale pouvait ainsi différer de toutes les évaluations initiales.
- Les patients concernés ont été informés, conformément à la réglementation en vigueur. Le comité scientifique et éthique de la cohorte LUCC a approuvé l'étude avant son démarrage.
- L'IA de Lifen provient du fine-tuning d'un modèle Mistral en open-source.

## Source

*Next-Generation Multicenter Studies: Using Artificial Intelligence to Automatically Process Unstructured Health Records of Patients with Lung Cancer across Multiple Institutions*  
*Annals of Oncology*, publication en ligne le 15 décembre 2025  
DOI: <https://doi.org/10.1016/j.annonc.2025.12.006>

## À propos de Gustave Roussy

Classé premier centre français, premier européen et sixième au niveau mondial, Gustave Roussy constitue un pôle d'expertise globale entièrement dédié aux patients vivant avec un cancer. L'Institut est un pilier fondateur du biocluster en oncologie Paris-Saclay Cancer Cluster. Source d'innovations thérapeutiques et d'avancées diagnostiques, l'Institut accueille chaque année près de 50 000 patients dont 3 500 enfants et adolescents et développe une approche intégrée entre recherche, soins et enseignement. Expert des cancers rares et des tumeurs complexes, Gustave Roussy traite tous les cancers, à tous les âges de la vie. Il propose à ses patients une prise en charge personnalisée qui allie innovation et humanité, où sont pris en compte le soin mais aussi la qualité de vie physique, psychologique et sociale. Avec 4 100 salariés répartis sur deux sites, Villejuif et Chevilly-Larue, Gustave Roussy réunit les expertises indispensables à une recherche de haut niveau en cancérologie ; 32 % des patients traités sont inclus dans des études cliniques.

Pour en savoir plus sur Gustave Roussy et suivre les actualités de l'Institut : [www.gustaveroussy.fr](http://www.gustaveroussy.fr), [X](#), [Facebook](#), [LinkedIn](#), [Instagram](#) et [Bluesky](#).

## À propos de Lifen

Fondée en 2015, Lifen est une entreprise spécialisée dans les données de santé qui s'est imposée comme un leader du numérique en santé. Experte en intelligence artificielle, Lifen offre une suite de solutions complète qui facilitent l'accès, la structuration et le partage sécurisé des données de santé dans le soin et dans la recherche. Aujourd'hui, plus de 800 hôpitaux et 20 000 professionnels de santé libéraux utilisent ses solutions pour réduire la charge administrative, personnaliser le soin et accélérer la recherche clinique. Lifen compte 160 collaborateurs basés partout en France. En recherche clinique, l'approche de Lifen a permis de lancer trois cohortes multicentriques avec 28 établissements partenaires en moins d'un an. En savoir plus sur Lifen: [www.lifen.fr](http://www.lifen.fr)

## Contacts presse

Lifen : Claire Guerin - [claire@katchreyners.com](mailto:claire@katchreyners.com) - 06 51 48 26 54

Gustave Roussy : Claire Parisel et Léona Pinto – [presse@gustaveroussy.fr](mailto:presse@gustaveroussy.fr) – 01 42 11 50 59 – 01 42 11 63 59 – 06 17 66 00 26