Institut de Cancérologie de l'Ouest

unicancer PAYS DE LA LOIRE

Alternance (Bac +5) - Data Science

Début : à partir de septembre 2026

Service Data Factory & Analytics – Direction de la Recherche et des Data - Site St Herblain

Date : 26/03/2026

CONTEXTE

L'Institut de Cancérologie de l'Ouest (ICO) Nantes-Angers est un centre d'excellence dans la prise en charge du cancer et la recherche. Labelisé Comprehensive Cancer Center (CCC) par l'Organisation of European Cancer Institutes (OECI), l'ICO est le 1er centre de lutte contre le cancer de province en France et le 3ème au niveau national après l'Institut Gustave Roussy (Villejuif) et l'Institut Curie (Paris) en termes de file active de patients et d'inclusion dans les essais cliniques.

Depuis 5 ans, l'ICO a placé les données de vie réelle (Real-World Data / RWD) et l'IA au centre de sa stratégie de recherche et d'innovation. Son service Data Factory & Analytics est une structure pluridisciplinaire, intégrée, dédiée à la collecte, à la qualification, à l'analyse et à la valorisation des données générées à l'occasion du soin de nos patients. Le service assure la maîtrise complète du cycle de vie des données depuis leur production lors de la prise en charge des patients au sein de l'ICO jusqu'à leur analyse et leur valorisation. Afin de développer son activité de recherche sur données de vie réelles, l'ICO développe son propre Entrepôt de Données de Santé (EDS). L'objectif est d'utiliser les différentes sources de données existantes à l'ICO dans le cadre de la recherche ou du soin afin de créer une unique base de données structurées contenant les variables considérées comme les plus importantes pour mener des travaux de recherche sur données observationnelles.

Aujourd'hui l'EDS est alimenté à partir des bases de données structurées disponibles à l'ICO et des travaux sont en cours dans le but d'extraire des données structurées à partir des documents des patients (comptes rendus de consultation, comptes rendus d'anatomopathologie, etc.). Depuis quelques années, de nombreuses études ont montré qu'il est possible d'extraire des données structurées à partir des comptes rendus en utilisant le NLP (Natural Language Processing) mais toutes ces études reposent sur une méthode nécessitant une longue et coûteuse phase d'annotation afin d'entrainer le modèle (1–6). A l'ICO nous avons réalisé un premier travail montrant les capacités d'un algorithme basé sur le Large Language Model (LLM) Mistral Large à extraire les données de 3 biomarqueurs du cancer du sein à partir des comptes-rendus d'anatomopathologie (7). D'autres travaux sont en cours afin d'extraire les dates de métastases et de progressions. L'alternance consistera à poursuivre les développements autour de cette pipeline LLM afin extraire d'autres données structurées à partir des comptes-rendus médicaux des patients.

MISSION

Poste rattaché au Service Data Factory & Analytics (Direction de la recherche et des data).

L'objectif principal est de développer une solution permettant d'automatiser le processus d'extraction d'informations pertinentes (périmètre des variables encore à définir) à partir de documents médicaux non structurés et d'évaluer les performances de cette solution.

Tâches principales :

- Compréhension des données médicales : familiarisation avec les différents types de comptes rendus médicaux. Analyse des spécificités linguistiques et des structures de ces documents.
- Appréhension de la pipeline d'extraction existante.

- Adaptation de la pipeline d'extraction existante et/ou développement d'une nouvelle : conception et mise en œuvre d'un pipeline automatisé utilisant Mistral AI pour extraire les variables d'intérêts à partir des documents médicaux, et permettant d'alimenter une base de données structurée.
- Évaluation de la performance de la solution en termes de précision, de rappel et de F1-score en utilisant une base de données manuellement saisie comme Gold Standard.
- Identification des opportunités d'amélioration et itération du modèle pour une extraction plus performante.
- Adaptation du process pour extraire différentes variables.

Cette alternance offre une opportunité unique d'acquérir des compétences pratiques en data science appliquée à la santé, tout en contribuant au développement d'une solution innovante essentielle pour exploiter des données médicales non structurées. L'alternant(e) travaillera en étroite collaboration avec une équipe multidisciplinaire composée de spécialistes en biostatistique, data science et en oncologie.

Références :

1. Schiappa R, Contu S, Culie D, Thamphya B, Chateau Y, Gal J, et al. RUBY: Natural Language Processing of French Electronic Medical Records for Breast Cancer Research. JCO Clin Cancer Inform. 2022 Jul;6:e2100199. doi:10.1200/CCI.21.00199 PubMed PMID: 35960900; PubMed Central PMCID: PMC9470144.
2. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc JAMIA. 2008;15(1):25–8. doi:10.1197/jamia.M2437 PubMed PMID: 17947622; PubMed Central PMCID: PMC2274870.
3. Holmes B, Chitale D, Loving J, Tran M, Subramanian V, Berry A, et al. Customizable Natural Language Processing Biomarker Extraction Tool. JCO Clin Cancer Inform. 2021 Aug;5:833–41. doi:10.1200/CCI.21.00017 PubMed PMID: 34406803.
4. Hanauer DA, Barnholtz-Sloan JS, Beno MF, Del Fiol G, Durbin EB, Gologorskaya O, et al. Electronic Medical Record Search Engine (EMERSE): An Information Retrieval Tool for Supporting Cancer Research. JCO Clin Cancer Inform. 2020 May;4:454–63. doi:10.1200/CCI.19.00134 PubMed PMID: 32412846; PubMed Central PMCID: PMC7265780.
5. Carrell DS, Halgrim S, Tran DT, Buist DSM, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol. 2014 Mar 15;179(6):749–58. doi:10.1093/aje/kwt441 PubMed PMID: 24488511; PubMed Central PMCID: PMC3939853.
6. Banerjee I, Bozkurt S, Caswell-Jin JL, Kurian AW, Rubin DL. Natural Language Processing Approaches to Detect the Timeline of Metastatic Recurrence of Breast Cancer. JCO Clin Cancer Inform. 2019 Oct;3:1–12. doi:10.1200/CCI.19.00034 PubMed PMID: 31584836.
7. Joseph E, Vallee P, Perennec T, Wagneur N, Frenel JS, Campone M, et al. Development and Assessment of a Pipeline for Extracting Structured Data From Free-Text Medical Reports Using a Large Language Model. JCO Clin Cancer Inform. 2026 Feb;10:e2500133. doi:10.1200/CCI-25-00133 PubMed PMID: 41707099; PubMed Central PMCID: PMC12928813.

PROFIL ATTENDU

En prévision de votre dernière année d'études (Bac +5) en Data Science, vous recherchez pour la rentrée prochaine une alternance. Vous devrez disposer de bonnes connaissances des modèles de traitement du langage et du machine learning et être force de proposition. Vous devez être à l'aise avec les langages de

programmation Python et/ou R et avoir une appétence pour les applications en santé et l'oncologie. De bonnes capacités de communication, orales et écrites, sont souhaitées.

Lieu de stage : Institut de Cancérologie de l'Ouest (ICO) - Site de Nantes / Saint-Herblain - Bd Professeur Jacques Monod, 44800 Saint-Herblain
Encadrant : Florent Le Borgne, Data Analyst - Statisticien
Date de début : à partir de septembre 2026
Durée : un an

Merci d'adresser votre CV et lettre de motivation à Florent Le Borgne
florent.leborgne@ico.unicancer.fr