

Offre de stage : Equité et performance prédictive du ciblage des campagnes de prévention dans le SNDS : quel apport d'un modèle d'IA prédictif par rapport à des règles métiers ?

La disponibilité accrue des données de santé pourrait permettre de mieux bâtir des stratégies territoriales et sociales de prévention, et de mieux concentrer les efforts sur les personnes les plus à risques (Cours des comptes, 2021). Une prévention personnalisée, s'appuyant sur des données telles que celles du Système National des Données de Santé (SNDS) pour construire sur des scores de risques construits à partir des parcours de soin de patients similaires, pourrait demain voir le jour¹. Historiquement, les campagnes de prévention nationale de l'assurance maladie sont le plus souvent ciblées sur des critères catégoriels simples (par exemple les femmes entre 50 et 74 ans pour l'invitation au dépistage organisé du cancer du sein), et peuvent aussi cibler des personnes à partir des données de prise en charge disponibles dans le SNDS (par exemple pour le dispositif Sophia à destination des personnes diabétiques chroniques).

Si le ciblage devient plus fin afin de mieux repérer les plus à risques et s'appuie sur des modèles d'IA prédictive, la question de l'équité de l'algorithme de ciblage se posera beaucoup plus clairement.

L'objet de ce stage est de comparer, sur le cas d'une campagne de dépistage existante basée sur des critères catégoriels (tel le diabète ou les dépistages organisés des cancers), le ciblage sur « règles métiers » et d'un ciblage à partir d'un modèle d'IA prédictive, tous deux basés sur les données du Système National des Données de Santé. Il s'agira d'évaluer :

- L'apport du modèle IA par rapport à la simple règle métier en termes de capacité à prédire un événement de santé adverse, et de significativement mieux cibler, ou non, les plus à risques.
- L'équité relative des règles de ciblage. Plusieurs critères de « non-discrimination statistique » peuvent être retenus : en fonction des caractéristiques socio-économiques des personnes, comment varie le taux de ciblage par groupe (l'effort initial de ciblage est-il équitablement réparti) ? les différences observées sont-elles justifiables par des états de santé différents, ou d'autres facteurs, comme les comportements de recours ou des inégalités d'accès, sont-ils prépondérants ? la capacité à repérer en amont est-elle équivalente en fonction des groupes ?
- Peut-on recalibrer le modèle afin d'obtenir des taux de recours au dépistage équivalents entre les groupes socio-économiques ?

On supposera que si les règles de ciblage changent (métiers > IA), les comportements de recours au dépistage restent identiques conditionnellement aux caractéristiques des individus, et peuvent donc être estimés à partir des comportements de recours observés pour les campagnes existantes basés sur des règles métiers. Dans le cadre du stage, il s'agira

¹ Depuis mai 2024, dans l'espace numérique de santé (ENS), des campagnes de prévention personnalisées et automatisées peuvent être mise en œuvre sur la base des données disponibles (auto-questionnaire patient principalement, mais aussi données de remboursement de l'assurance maladie présents dans l'ENS).

d'évaluer la capacité à effectivement amener une population donnée à se faire dépister, et non pas l'efficacité de la campagne de dépistage à éviter des événements de santé adverse, ce qui supposerait d'établir pour une situation contrefactuelle sans campagne de dépistage.

Ce stage s'inscrit dans un projet de recherche SeqNDS, conjoint Drees-Inria. En s'appuyant sur la richesse des données disponibles, ce projet ambitionne de développer des modèles capables de prévoir des événements dans la suite des parcours de soin du patient, en particulier les hospitalisations avec un diagnostic incident. Le modèle d'IA issu de ces travaux sera considéré comme une donnée par le stagiaire. Il sera appliqué sur un échantillon « test » de la population à 4%, pour lequel les données du SNDS ont été chaînées à des caractéristiques socio-économiques des personnes (EDP-Santé), pour définir les personnes qui pourraient être ciblées pour un dépistage une année (ou qui l'ont été, et quel a été leur recours pour la campagne de dépistage classique), et la probabilité qu'elle le développe effectivement la pathologie à éviter l'année suivante.

Profil : Étudiant M1/M2 avec de solides compétences en Python, statistiques ainsi que de bonne compétences en statistiques ou machine learning.

Bibliographie indicative :

- [BEHRT](#)
- [MED-BERT](#)
- [What is the Bureaucratic Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy | Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency](#)
- <https://fairmlbook.org/>