Offre n°2025-09132

Doctorant F/H Apprentissage statistique causal pour évaluer l'impact des interventions sur les trajectoires du diabète en utilisant les dossiers patient électroniques

Type de contrat : CDD

Niveau de diplôme exigé : Bac + 5 ou équivalent

Fonction : Doctorant

Niveau d'expérience souhaité : Jeune diplômé

A propos du centre ou de la direction fonctionnelle

Le centre de recherche Inria de Saclay a été créé en 2008. Sa dynamique s'inscrit dans le développement du plateau de Saclay, en partenariat étroit d'une part avec le pôle de l'**Université Paris-Saclay** et d'autre part avec le pôle de l'**Institut Polytechnique de Paris** . Afin de construire une politique de site ambitieuse, le centre Inria de Saclay a signé en 2021 des accords stratégiques avec ces deux partenaires territoriaux privilégiés.

Le centre compte **40 équipes-projets**, dont 32 sont communes avec l'Université Paris-Saclay ou l'Institut Polytechnique de Paris. Son action mobilise **plus de 600 personnes**, scientifiques et personnels d'appui à la recherche et à l'innovation, issues de 54 nationalités.

Le centre Inria Saclay - Île-de-France est un acteur essentiel de la recherche en sciences du numérique sur le plateau de Saclay. Il porte les valeurs et les projets qui font l'originalité d'Inria dans le paysage de la recherche : l'excellence scientifique, le transfert technologique, les partenariats pluridisciplinaires avec des établissements aux compétences complémentaires aux nôtres, afin de maximiser l'impact scientifique, économique et sociétal d'Inria.

Contexte et atouts du poste

Les dossiers patient électroniques permettent d'avoir une vision sans précédent sur les trajectoires d'évolution des patients diabétiques en s'affranchissant des biais des cohortes des études cliniques : grand nombre de patients, non-exclusion des situations complexes, suivi pluri-annuel. Au-delà des modèles prédictifs du futur de la maladie, on s'intéresse aussi à l'effet de différentes interventions possibles sur le devenir du patient, afin de pouvoir personnaliser la prise en charge du

diabète pour éviter les complications : c'est le cadre de l'inférence causale.

Il existe de nombreuses méthodes pour analyser l'effet de traitement individualisé pour des variables de type binaire (survenue d'une complication à un horizon donnné) ou continu (valeur de la glycémie), cependant d'autres variables d'intérêt nécessitent une analyse de survie (survenue d'une complication sans horizon fixé) en raison du phénomène de censure de certaines observations, par exemple si un patient déménage. De nouveaux développements métholodogiques sont donc nécessaires pour mesurer la capacité des options thérapeutiques à retarder voire empêcher la survenue de complications du diabète, sources de pathologies chroniques dont la prise en charge est très lourde. Les approches d'apprentissage statistique proposées dans le cadre de ce travail de thèse devront également prendre en compte d'autres aspects spécifiques des dossiers patient électroniques : présence de valeurs manquantes informatives, sources de données hétérogènes, biais temporels.

Context
Electronic health records, such as hospital-level databases of routine care, gather a large amount of health data across many individuals. Using them to improve health requires adapting interventions, which calls for causal analyses. The very rich data in electronic health records enables individualizing decisions, but this requires powerful models to adapt to individuals as well as to the complexity of the data (comprising clinical notes, irregular sequences...), typically using machine learning for causal inference [1].

Another challenge of this data is that it has an important time-wise component. Consequently, for a given time window (typically that for which the intervention scenario of interest is studied), the outcome or the intervention is censored (missing). Analysis methods, including machine-learning models, must then be corrected for this censoring, for instance, with corresponding inverse probabilities [2]. Correctly designing a study without time-related biases is challenging [1], and there is a lack of tools that both help such causal analysis on time-wise data and compute individualized effects.

Chronic diseases are specific health burdens that could particularly benefit from the good use of already-collected routine-care data. Indeed, the corresponding patients interact often and over a long time with the health system, leading to rich data. The stakes are high when health interventions exist that can improve the health outcomes of patients, for instance, those with diabetes, where the stakes are to avoid complications.

Goals
The goal of this project is to develop estimators of heterogeneous causal effects (CATE) in the presence of censoring. We will consider two possible strategies:
1) adapting existing estimators of the CATE [1] to censored data –as in [3], but using models adapted from [2] which perform best on health records–,
2) adapting the cloning, censoring, and weighting approach [4] to machine-learning estimators.

The techniques will be applied to a large cohort of 1 million diabetes patients that we have extracted and consolidated from the AP-HP health data mart. The questions of interest are: what are the markers of complications and the related beneficial interventions? While this information is already well known in the

medical literature when working with research-level data, the challenge is to find what in the routine-care data can drive better decisions.

Mission confiée

Outils d'apprentissage statistiques et leurs applications à la santé

machine learning

Les dossiers médicaux électroniques, tels que les bases de données hospitalières sur les soins de routine, rassemblent une grande quantité de données sur la santé de nombreux individus. Pour les utiliser afin d'améliorer la santé, il faut adapter les interventions, ce qui nécessite des analyses causales. Les données très riches des dossiers médicaux électroniques permettent d'individualiser les décisions, mais cela nécessite des modèles puissants pour s'adapter aux individus ainsi qu'à la complexité des données (comprenant des notes cliniques, des séquences irrégulières...), en utilisant généralement l'apprentissage automatique pour l'inférence causale [1].

Un autre défi de ces données est qu'elles ont une composante temporelle importante. Par conséquent, pour une fenêtre temporelle donnée (typiquement celle pour laquelle le scénario d'intervention d'intérêt est étudié), le résultat ou l'intervention est censuré (manquant). Les méthodes d'analyse, y compris les modèles d'apprentissage automatique, doivent alors être corrigées de cette censure, par exemple avec les probabilités inverses correspondantes [2]. Concevoir correctement une étude sans biais liés au temps est un défi [1], et il y a un manque d'outils qui aident à la fois une telle analyse causale sur des données temporelles et qui calculent des effets individualisés.

Les maladies chroniques sont des fardeaux sanitaires spécifiques qui pourraient particulièrement bénéficier d'une bonne utilisation des données déjà collectées sur les soins de routine. En effet, les patients concernés interagissent souvent et longtemps avec le système de santé, ce qui permet d'obtenir des données très riches. Les enjeux sont importants lorsqu'il existe des interventions sanitaires susceptibles d'améliorer l'état de santé des patients, par exemple ceux atteints de diabète, pour lesquels il s'agit d'éviter les complications.

Principales activités

Le but de ce projet est de développer des estimateurs d'effets causaux hétérogènes (CATE) en présence de censure. Nous envisagerons deux stratégies possibles :
1) adapter les estimateurs existants du CATE [1] aux données censurées - comme dans [3], mais en utilisant des modèles adaptés de [2] qui donnent les meilleurs résultats sur les dossiers médicaux -,
2) l'adaptation de l'approche du clonage, de la censure et de la pondération [4] aux estimateurs par apprentissage automatique.

Les techniques seront appliquées à une grande cohorte d'un million de patients diabétiques que nous avons extraite et consolidée à partir des données de santé de l'AP-HP. Les questions qui nous intéressent sont les suivantes : quels sont les marqueurs de complications et les interventions bénéfiques correspondantes ?

Alors que ces informations sont déjà bien connues dans la littérature médicale lorsque l'on travaille avec des données de niveau recherche, le défi consiste à trouver ce qui, dans les données de soins de routine, peut conduire à de meilleures décisions.

Compétences

- Bonne formation en statistique, idéalement avec des connaissances en biostatistique.
- Connaissance de l'apprentissage automatique
- Maîtrise raisonnable du français
- Maîtrise de Python, pandas, scikit-learn.
- Intérêt marqué pour les problèmes de santé
- Esprit curieux.
- Good statistical background, ideally with biostatistical knowledge.
- Machine learning background
- Reasonable proficiency in French
- Proficiency in Python, pandas, scikit-learn.
- A strong interest for health problems
- Curious mindset.

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Rémunération

2200€ brut/mois

Informations générales

- **Thème/Domaine :** Neurosciences et médecine numériques
  Statistiques (Big data) (BAP E)
- **Ville :** Palaiseau
- **Centre Inria :** Centre Inria de Saclay
- **Date de prise de fonction souhaitée :** 2025-11-01
- **Durée de contrat :** 3 ans

- **Date limite pour postuler :** 2025-10-31

Contacts

- **Équipe Inria :** SODA
- **Directeur de thèse :**
  Abecassis Judith / judith.abecassis@inria.fr

A propos d'Inria

Inria est l’institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l’interface d’autres disciplines. L’institut fait appel à de nombreux talents dans plus d’une quarantaine de métiers différents. 900 personnels d’appui à la recherche et à l’innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

- Bonne formation en statistique, idéalement avec des connaissances en biostatistique.
- Connaissance de l'apprentissage automatique
- Maîtrise raisonnable du français
- Maîtrise de Python, pandas, scikit-learn.
- Intérêt marqué pour les problèmes de santé
- Esprit curieux.
- Good statistical background, ideally with biostatistical knowledge.
- Machine learning background
- Reasonable proficiency in French
- Proficiency in Python, pandas, scikit-learn.
- A strong interest for health problems
- Curious mindset.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du

recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.