Offre de stage de Master 2 en Biostatistique

Construction de scores prédictifs de facteurs liés au mode de vie à partir du Système National des Données de Santé dans la cohorte Constances

Contexte

Un des objectifs de l'équipe Biostatistique en grande dimension du Centre de recherche en épidémiologie et santé des populations (CESP, Inserm U1018) est de développer des méthodes statistiques pour l'exploitation de bases de données massives et complexes, comme celles du Système National des Données de Santé (SNDS), afin d'évaluer les effets des médicaments en population.

Le SNDS contient des informations détaillées sur les remboursements de consommation de soins et les séjours hospitaliers, pour la quasi-totalité de la population française.
Néanmoins, l'exploitation de ces données à des fins de recherches pose certains défis méthodologiques. En particulier, des informations relatives à certains facteurs d'ajustement couramment utilisés dans les études épidémiologiques ne sont pas disponibles, pouvant mener à des biais importants dans les analyses.

La cohorte Constances est une cohorte épidémiologique constituée d'un échantillon représentatif d'environ 200 000 adultes âgés de 18 à 69 ans à l'inclusion pour lesquels sont recueillis, via des examens médicaux et questionnaires, un très grand nombre d'indicateurs de santé et de facteurs qui leur sont potentiellement associés. Les données recueillies pour cette cohorte sont par ailleurs chainées aux données du SNDS.

Objectif

A partir d'une extraction des données de Constances, chaînées aux données du SNDS, l'objectif du stage est de développer et d'évaluer les performances de scores prédictifs construits à partir des données disponibles dans le SNDS pour plusieurs facteurs d'ajustement mesurés dans les questionnaires de la cohorte Constances. Les méthodes envisagées seront appropriées au contexte de la grande dimension (très grand nombre de prédicteurs), telles que les régressions pénalisées de type lasso ou les arbres de classification.

Ces scores seront développés pour plusieurs facteurs de confusion usuellement rencontrés dans les études épidémiologiques : le statut tabagique, l'activité physique ou encore l'indice de masse corporelle. Les performances du score développé pourront être comparées à celles d'autres algorithmes proposés dans la littérature, en particulier pour l'indicateur de statut tabagique (Lassalle et al. 2021; Faust et al. 2023). A terme, il sera possible d'utiliser ces scores comme variable d'ajustement dans des analyses au sein du SNDS.

A partir des données de la cohorte Constances, une comparaison sera faite entre l'effet de l'ajustement sur les variables originales (statut tabagique, activité physique, indice de masse corporelle) ou sur les scores sur la relation entre deux variables, afin d'estimer la confusion résiduelle après ajustement sur les scores.

Profil recherché

- M2 avec une forte composante en biostatistique et en apprentissage statistique
- Bonne maîtrise du langage R indispensable

Missions confiées au stagiaire

- Prise en main des données de la cohortes Constances (questionnaires et SNDS)
- Mise en œuvre des analyses statistiques pour la construction de scores prédictifs du statut tabagique, de l'activité physique et de l'IMC.
- Rédaction d'un rapport de stage selon le plan d'un article scientifique en vue d'une publication.

Équipe d'accueil

Équipe Biostatistique en grande dimension pour la sécurité des médicaments et la génomique
Centre de Recherche en Epidémiologie et Santé des Population (CESP) – INSERM U1018

Lieu du stage

Bâtiment INSERM de l'hôpital Paul Brousse, 16 avenue Paul Vaillant-Couturier, Villejuif

Durée du stage

6 mois

Encadrement

Émeline Courtois, Ismaïl Ahmed, Anne Thiébaut, Alexis Elbaz

Coordonnées

CV et lettre de motivation sont à adresser à emeline.courtois@inserm.fr; ismail.ahmed@inserm.fr et anne.thiebaut@inserm.fr

Références bibliographiques

Faust, Irene, Mark Warden, Alejandra Camacho-Soto, Brad A. Racette, et Susan Searles Nielsen. 2023. « A predictive algorithm to identify ever smoking in medical claims-based epidemiologic studies ». *Annals of Epidemiology* 85 (septembre): 59-67.e6. https://doi.org/10.1016/j.annepidem.2023.04.019.

Lassalle, M., T. Le Tri, P. Afchain, M. Camus, J. Kirchgesner, M. Zureik, et R. Dray-Spira. 2021. « Use of Proton Pump Inhibitors and Risk of Pancreatic Cancer: A Nationwide Case-Control Study Based on the French National Health Data System (SNDS). » *Cancer Epidemiol Biomarkers Prev*, décembre. https://doi.org/10.1158/1055-9965.EPI-21-0786.