Offre de stage de Master 2 en Biostatistique – Janvier 2027

Étude de la variabilité des probabilités prédites en grande dimension et impact sur l'ajustement de modèles épidémiologiques

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 pour faciliter la recherche biomédicale. Ainsi, un des projets actuels de l'équipe[1] vise à développer des modèles prédictifs de facteurs liés au mode de vie (par exemple : le tabagisme, la consommation d'alcool, etc. sous forme de facteurs binaires) qui sont souvent absents dans ces bases de données, mais qui sont des facteurs d'ajustements cruciaux dans de nombreuses études épidémiologiques.

Pour une réponse d'intérêt binaire, les performances prédictives d'un modèle s'évaluent selon deux dimensions :

- la discrimination : soit la capacité du modèle à classer correctement les individus, mesurée par exemple par l'Area Under Curve (AUC) ;
- la calibration : soit la concordance entre les résultats observés et les prédictions obtenues, mesurée par exemple à l'aide d'une courbe de calibration (Van Calster et al., 2024).

Des travaux récents de Yusufujiang et al. (Yusufujiang et al., 2026) ont montré que des modèles issus de méthodes d'apprentissages statistiques différentes peuvent présenter des performances prédictives globalement équivalentes en termes de discrimination (valeurs d'AUC proches) tout en attribuant des probabilités individuelles très différentes aux mêmes individus. Cette étude a été menée dans un contexte où le nombre de variables explicatives était très limité.

L'objectif de ce stage est d'explorer, à l'aide de simulations, ce phénomène dans un contexte de grande dimension, caractérisé par un nombre élevé de variables. Par ailleurs, ce travail visera à évaluer l'impact de ces divergences de distribution des scores prédits lorsque ces scores sont utilisés comme facteurs d'ajustement dans des analyses épidémiologiques subséquentes.

Objectif et méthodes

En s'appuyant sur des données réelles issues du chaînage du SNDS avec la cohorte Constances, l'étudiant·e mettra en œuvre une technique de simulations conservant les caractéristiques complexes du jeu de données réelles considéré, en s'inspirant du modèle de simulation proposé par (Austin et al., 2024).

[1] Le projet SCOP : Construction de scores prédictifs de facteurs liés au mode de vie, chaînage de la cohorte Constances au Système National des Données de Santé.

L'étudiant·e développera des modèles prédictifs des réponses d'intérêt simulées à l'aide de méthodes d'apprentissages statistiques appropriées au contexte de la grande dimension, comme la régression Lasso (Tibshirani, 1996) ou les arbres de classification boostés (Hastie et al., 2009).

Le travail s'articulera autour de deux axes principaux :

- Analyse de la divergence des scores : à travers différents scénarios de simulation, l'étudiant·e cherchera à identifier quand et pourquoi différentes méthodes produisent des probabilités très différentes, et comment cela impacte les métriques de discrimination et de calibration.
- Impact épidémiologique : l'étudiant·e étudiera, via des simulations, l'influence de la distribution et des performances des scores sur l'estimation d'une relation épidémiologique lorsque ces scores sont utilisés comme facteurs d'ajustement.

L'objectif final est de quantifier l'instabilité des prédictions individuelles en grande dimension et d'alerter sur les biais potentiels induits lors de l'utilisation de ces scores dans des analyses subséquentes.

Profil recherché

- Solide bagage en mathématiques : formation initiale en mathématique/statistique souhaitée
- M2 avec une forte composante en biostatistique et en apprentissage statistique
- Bonne maîtrise du langage R indispensable
- Bonne maîtrise de l'anglais (écrit et oral)

Missions confiées au stagiaire

- Présentation orale de son travail au sein des séminaires d'équipe ou d'unité
- Rédaction d'un rapport de stage selon le plan d'un article scientifique en vue d'une publication.

Encadrement et équipe d'accueil

Ce projet sera supervisé par Émeline Courtois (Chaire de Professeur Junior), Ismaïl Ahmed (Chargé de Recherche) de l'équipe Biostatistique en grande dimension pour la pharmacoépidémiologie et la génomique, du Centre de Recherche en Epidémiologie et Santé des Population (CESP) – INSERM UMR1018

Lieu du stage

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

Durée du stage

6 mois, à pourvoir dès que possible à partir de janvier 2027.

Coordonnées

CV, lettre de motivation et relevés de notes de licence et master (ou équivalent) sont à adresser à emeline.courtois@inserm.fr et ismail.ahmed@inserm.fr.

Références

Austin, P. C., Lee, D. S., & Wang, B. (2024). The relative data hungriness of unpenalized and penalized logistic regression and ensemble-based machine learning methods : The case of calibration. *Diagnostic and Prognostic Research*, *8*(1), 15. https://doi.org/10.1186/s41512-024-00179-z

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Additive Models, Trees, and Related Methods. In T. Hastie, R. Tibshirani, & J. Friedman (Éds.), *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (p. 295-336). Springer. https://doi.org/10.1007/978-0-387-84858-7_9

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *58*(1), 267-288. https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G. M., Hernandez-boussard, T., Timmerman, D., Mclernon, D. J., Smeden, M. V., & Steyerberg, E. W. (2024). *Performance evaluation of predictive AI models to support medical decisions : Overview and guidance* (arXiv:2412.10288). arXiv. https://doi.org/10.48550/arXiv.2412.10288

Yusufujiang, M., Navarro, C. L. A., Damen, J. A., Takada, T., Geersing, G.-J., Hooft, L., Schuit, E., Moons, K. G., De Jong, V. M., & Van Smeden, M. (2026). Prediction models developed using artificial intelligence : Similar predictive performances with highly varying predictions for individuals – an illustration in deep vein thrombosis. *Diagnostic and Prognostic Research*, *10*(1), 1. https://doi.org/10.1186/s41512-025-00216-5