

# COMPLEMENTS A JOINDRE

## AU DEPOT DU SUJET DE THESE

### INTITULE DE LA THESE

---

Développement et évaluation de scores prédictifs de facteurs liés au mode de vie à partir de méthodes d'apprentissage statistique : application au Système National des Données de Santé

### PROBLEMATIQUE SCIENTIFIQUE GENERALE

---

Le Système National des données de Santé (SNDS) est une large base de données médico-administratives qui regroupe les informations individuelles et anonymisées sur la consommation de soins de la quasi-totalité de la population française. Malgré son exhaustivité et sa taille, le SNDS ne dispose pas d'informations renseignant directement sur les facteurs liés au mode de vie. Or, les caractéristiques du mode de vie sont bien souvent des facteurs de confusion pour de nombreux événements de santé et sont, à ce titre, des facteurs d'ajustement incontournables dans de nombreuses études épidémiologiques. A défaut, les études pharmacoépidémiologiques menées à partir du SNDS ont recours à des algorithmes, construits à partir des données disponibles, telles que les remboursements de médicaments, les diagnostics d'hospitalisation ou encore les affections de longue durée. Ces indicateurs reposent généralement sur des règles de décision issues de connaissance d'experts.

De par la richesse et le volume des informations disponibles dans le SNDS, cette base de données se prête potentiellement au développement de modèles prédictifs basés sur des méthodes d'apprentissage automatique adaptées au contexte de la grande dimension. Néanmoins, pour ce faire, il est nécessaire de pouvoir disposer pour une partie de la population de l'information à prédire afin de pouvoir entraîner ces modèles de prédiction. Dans ce contexte, les grandes cohortes en population générale, chaînée au SNDS telles que la cohorte Constances, représentent une ressource particulièrement précieuse.

La construction classique d'un modèle prédictif repose sur la séparation du jeu de données en deux sous-ensembles : un ensemble d'apprentissage, sur lequel le modèle est construit, et un ensemble de validation, sur lequel les performances prédictives du modèle sont évaluées. En comparant les prédictions obtenues et les réponses observées sur le jeu de validation, on peut estimer différentes métriques (Van Calster et al. 2024) qui permettent d'évaluer les performances prédictives du modèle. Dans le cas d'une réponse d'intérêt binaire, les performances prédictives sont décrites selon deux dimensions : la discrimination (capacité du modèle à classer correctement les individus) et la calibration (écart entre les résultats observés et les prédictions obtenues). Cependant, il a été montré depuis longtemps que cette approche peut être sous-optimale voire trompeuse (Steyerberg et al. 2001). En effet, elle conduit à se priver d'une partie importante des données pour l'étape d'apprentissage et/ou à estimer les performances prédictives sur un jeu de données trop petit, entraînant ainsi une grande variabilité dans les estimations fournies.

Des alternatives plus robustes consistent à entraîner le modèle sur l'ensemble des données disponibles, puis à estimer une mesure d'optimisme afin d'évaluer les performances prédictives réelles du modèle. L'optimisme est défini comme l'écart entre les performances

apparentes, déterminées sur l'échantillon ayant servi au développement du modèle – par construction « optimistes », et les performances réelles attendues dans la population sous-jacente. Cet écart est estimé à l'aide du bootstrap (Harrell et al. 1996). Une autre méthode de rééchantillonnage, la validation croisée k-folds, consiste également à utiliser toutes les données disponibles pour développer le modèle. L'ensemble d'un jeu de données est partitionné en k sous-échantillons de sorte que chaque ensemble serve exactement k-1 fois d'ensemble d'apprentissage et une fois d'ensemble de validation. La performance moyenne sur les k itérations est considérée comme une estimation de la performance du modèle final. Dans une publication récente qui visait à comparer ces différentes approches, la méthode de bootstrap pour le calcul de l'optimisme s'est avérée la plus fiable pour obtenir une estimation des performances prédictives (Collins et al. 2024) .

Néanmoins, l'utilisation de ces méthodes d'évaluation appliquées à des méthodes d'apprentissage statistique, où la construction des modèles repose sur l'optimisation d'un ou plusieurs hyperparamètres (ex : paramètre de régularisation pour une régression pénalisée, taux d'apprentissage ou profondeur des arbres pour des arbres de classification boostés), restent peu étudiées.

Dans le cadre du projet SCOP<sup>1</sup> financé par le Health Data Hub, nous avons cherché à développer un modèle prédictif du statut tabagique binaire (ever/never) à partir des données de la cohorte Constances. Nous avons été confrontés à la difficulté d'évaluer de manière fiable les performances prédictives de nos modèles développés dans le contexte de la grande dimension. En particulier, le calcul de l'optimisme a montré que les modèles ajustés sur les échantillons bootstrap étaient bien plus complexes que le modèle développé sur les données originales, ce qui pose question sur la pertinence de l'évaluation des performances prédictives de ce dernier avec cette approche. Afin de proposer une méthode fiable d'estimation de ces performances, il devient indispensable d'approfondir notre compréhension théorique de cette problématique d'évaluation.

## OBJECTIFS SCIENTIFIQUES DE LA THESE

---

- 1) Évaluation des méthodes d'estimation des performances prédictives de modèles adaptés au contexte de la grande dimension via une étude de simulations, et proposition d'une stratégie pour intégrer dans l'estimation de l'optimisme de nos modèles la variabilité liée à l'optimisation d'hyperparamètres.
- 2) Application au SNDS, construction et évaluation des scores prédictifs :
  - a. de la consommation d'alcool,
  - b. de la corpulence.
- 3) Évaluation de l'influence des performances prédictives des scores issus de l'apprentissage statistique sur le biais et la variance des estimations d'association dans les études observationnelles

---

<sup>1</sup> Construction de SCORes Prédictifs de facteurs liés au mode de vie : chaînage de la cohorte Constances au Système National des Données de Santé (<https://www.constances.fr/espace-scientifique/recherches-et-etudes/construction-de-scores-predictifs-de-facteurs-lies-au-mode-de-vie-chainage-de-la-cohorte-constances-au-systeme-national-des-donnees-de-sante/>)

## TRAVAUX PROJETES - METHODES & MOYENS

---

Le premier objectif de ce projet consiste à comparer différentes approches permettant de mesurer les performances prédictives de modèles construits à l'aide de régressions pénalisées (régression lasso, ridge). À cette fin, une étude de simulation approfondie sera menée. Afin de préserver les spécificités et la complexité des bases médico-administratives, nos simulations seront basées sur des données réelles. Les différents scénarios envisagés feront varier le nombre d'observations et de variables disponibles dans le jeu de données, la nature de la réponse d'intérêt (binaire, catégorielle ou continue) et la complexité du modèle de simulation considéré comme oracle. La contrainte computationnelle liée à l'utilisation du bootstrap dans le cadre d'une étude de simulation à grande échelle, sera explicitement prise en compte grâce à l'optimisation du code et la parallélisation des calculs. Notre objectif final est de proposer une méthode de calcul d'optimisme capable d'intégrer la variabilité due à l'optimisation du paramètre de régularisation, afin de « corriger » les différents indicateurs des performances prédictives de nos modèles. Par la suite, nous élargirons le cadre de cette étude aux méthodes ensemblistes comme les forêts aléatoires les arbres de classification boostés.

Le deuxième objectif de ce projet de thèse vise à développer, au sein du SNDS, des modèles prédictifs relatifs à deux facteurs liés au mode de vie : la consommation d'alcool et la corpulence. Ces deux variables, largement reconnues comme facteurs de risque pour de nombreuses maladies, ne sont pas directement renseignées dans le SNDS, ce qui pose des défis méthodologiques pour les études observationnelles conduites sur ces données. Le groupement d'intérêt EPI-Phare a proposé des indicateurs binaires construits à partir de diagnostics et de médicaments associés (Tran et al. 2025) : un indicateur de consommation excessive d'alcool, et un indicateur d'obésité. A notre connaissance ces indicateurs n'ont pas fait l'objet de validation dans le contexte des données françaises. En revanche, cet effort de validation a déjà été entrepris pour d'autres bases de données médico-administratives concernant l'obésité. Dans ce travail, nous développerons et évaluerons des scores prédictifs (le terme « score » désignant l'estimation fournie par le modèle) qui pourront éventuellement servir de variables d'ajustement dans les études menées sur le SNDS. Les réponses d'intérêt de nos modèles seront issues des données disponibles dans Constances, tandis que les variables explicatives seront exclusivement issues du SNDS. Une réflexion méthodologique sera menée afin de déterminer la meilleure façon de définir nos réponses d'intérêt à partir des informations disponibles pour caractériser au mieux ces comportements de santé. Différentes stratégies seront envisagées : binariser, catégoriser ces facteurs d'intérêts (ex : différentes catégories de consommateurs d'alcool, utilisation de la classification de l'Organisation Mondiale de la Santé pour l'Indice de Masse Corporel) ou les considérer comme des variables continues. A l'aide des résultats précédemment obtenus, les performances de nos modèles prédictifs pourront être évaluées de manière appropriée. En parallèle de ce travail, les indicateurs précédemment proposés seront implémentés, et nous serons en capacité d'évaluer leurs performances en termes de discrimination.

Dans un troisième temps, nous évaluerons l'impact du pouvoir prédictif du score lorsqu'il est considéré comme facteur d'ajustement dans l'analyse d'une association entre une exposition et une pathologie d'intérêt. En effet, considérer le score comme une variable d'ajustement « classique » pose plusieurs défis méthodologiques. D'une part, le score ne reflète pas parfaitement la relation entre la variable non observée (approchée par le score) et la pathologie, ce qui peut laisser subsister de la confusion résiduelle. D'autre part, la variabilité

inhérente à l'étape d'estimation du score est rarement prise en compte dans le modèle final, ce qui peut conduire à une sous-estimation de la variance de l'estimateur. Cette évaluation s'appuiera sur un plan de simulation analogue à celui développé dans le premier axe. La variable à prédire sera simulée pour être soit un facteur de confusion dans la relation entre l'exposition et la pathologie d'intérêt, soit un prédicteur de la pathologie d'intérêt. L'analyse des résultats permettra de caractériser les conditions sous lesquelles l'ajustement sur un score prédictif est approprié.

Ce projet de thèse s'inscrit dans la continuité du projet SCOP, tout en visant à approfondir les questionnements théoriques soulevés lors de sa mise en œuvre. L'accès aux données de la cohorte Constances, via la bulle sécurisée du Centre d'Accès Sécurisé aux Données (CASD), est financé jusqu'au premier trimestre 2027 par le projet en cours. Pour l'accès aux données sur la suite de la thèse, il pourra être prise en charge par les fonds de la Chaire de Professeur Junior d'Émeline Courtois. Après accord du Health Data Hub et de l'UMS 11, le doctorant sera intégré au projet ; son compte utilisateur sera créé à l'issue d'une formation obligatoire auprès du CASD.

## Références

- Collins, Gary S., Paula Dhiman, Jie Ma, et al. 2024. « Evaluation of Clinical Prediction Models (Part 1): From Development to External Validation ». *Research Methods & Reporting. BMJ* 384 (janvier): e074819. <https://doi.org/10.1136/bmj-2023-074819>.
- Harrell, Frank E., Kerry L. Lee, et Daniel B. Mark. 1996. « Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors ». *Statistics in Medicine* 15 (4): 361-87. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4<361::AID-SIM168>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4).
- Steyerberg, Ewout W., Frank E. Harrell, Gerard J. J. M. Borsboom, M. J. C. Eijkemans, Yvonne Vergouwe, et J. Dik F. Habbema. 2001. « Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis ». *Journal of Clinical Epidemiology* 54 (8): 774-81. [https://doi.org/10.1016/S0895-4356\(01\)00341-9](https://doi.org/10.1016/S0895-4356(01)00341-9).
- Tran, Anh, Mahmoud Zureik, Jeanne Sibiude, et al. 2025. « First-Trimester Exposure to Macrolides and Risk of Major Congenital Malformations Compared with Amoxicillin: A French Nationwide Cohort Study ». *PLOS Medicine* 22 (4): e1004576. <https://doi.org/10.1371/journal.pmed.1004576>.
- Van Calster, Ben, Gary S. Collins, Andrew J. Vickers, et al. 2024. « Performance Evaluation of Predictive AI Models to Support Medical Decisions: Overview and Guidance ». arXiv:2412.10288. Prépublication, arXiv, décembre 13. <https://doi.org/10.48550/arXiv.2412.10288>.

## CALENDRIER PREVISIONNEL DES TRAVAUX ET DES PUBLICATIONS

---

(La thèse doit représenter le volume de 3 articles signés en 1<sup>ère</sup> position par le doctorant, dont au moins 2 doivent être acceptés pour publication au moment de la soutenance)

Octobre 2026 – Décembre 2027

Revue de la littérature, élaboration du plan de simulation.

Implémentation des simulations et analyse des résultats avec des modèles basés sur des régressions pénalisées.

Réflexion sur une stratégie pour intégrer dans l'estimation de l'optimisme de nos modèles la variabilité liée à l'optimisation du paramètre de régularisation.

Extension aux méthodes ensemblistes.

Publication 1.

Janvier 2028 – Août 2028

Construction de scores prédictifs pour la consommation d'alcool et pour la corpulence, à partir des variables disponibles dans le SNDS.

Publication 2 et 3 : un article par score développé.

Septembre 2028 – Mai 2029

Reprise des scénarios de simulations développés dans l'axe 1.

Enrichissement avec la simulation d'une association épidémiologie d'intérêt.

Quantification de l'impact des performances prédictives d'un score utilisé comme facteur d'ajustement sur l'estimation de cette association.

Publication 4

Juin 2029-Octobre 2029

Rédaction du mémoire de thèse, préparation à la soutenance.