

Proposition de stage de fin d'étude :

Modèle Prédicatif des réhospitalisations chez des sujets hospitalisés pour une Insuffisance Cardiaque : une analyse des données PMSI par des méthodes de Machine Learning

Contexte médicale :

L'insuffisance cardiaque (IC) est une déficience de la fonction cardiaque qui ne permet plus d'assurer un débit cardiaque suffisant pour couvrir les besoins énergétiques de l'organisme. Les causes sont variées mais la principale est l'ischémie coronarienne. L'IC est très fréquente et touche 1,5 millions de personnes en France. Elle est responsable de 70 000 morts par an.

Cliniquement, on distingue 1) l'insuffisance cardiaque chronique, qui se manifeste par un essoufflement inhabituel à l'effort et 2) des décompensations aiguës qui nécessitent souvent une hospitalisation. Les décès surviennent essentiellement durant ces hospitalisations qui représentent également un coût considérable.

Afin de mieux comprendre et appréhender les facteurs prédictifs de la survenue d'hospitalisations pour une poussée d'insuffisance cardiaque, nous souhaitons réaliser une analyse de la base de données du PMSI. Le PMSI est une base de données regroupant l'ensemble des prises en charge à l'hôpital entre 2010 et 2020. Il contient notamment l'ensemble des codes diagnostiques (codage CIM-10) et actes médicaux (codage CCAM) correspondant aux séjours.

L'objectif clinique de ce stage est de trouver un modèle permettant de prédire les réhospitalisations dans un contexte de grande dimension.

Méthodologie :

Plusieurs modèles seront testés pour prédire les réhospitalisations. Il s'agit des modèles les plus classiques d'apprentissage supervisé et plus spécifiquement des classifications : régression logistique, plus proche voisin, forêt aléatoire, svm, réseau de neurones (perceptron). Différentes étapes seront menées successivement pendant ce stage :

1. Utilisation des modèles dans leur paramétrage de base
2. Optimisation des paramètres de ces modèles par cross-validation
3. Combinaison des modèles précédents avec des modèles de réduction de dimension comme les méthodes Lasso ou Elastic Net.

Chacun des modèles sera évalué selon la sensibilité, la spécificité et l'accuracy. L'importance des variables sera étudiée pour tous les modèles afin d'identifier des facteurs prédictifs.

En fonction de l'avancé du stage deux pistes de travail seront envisagées :

- Etude des modèles précédents sur des données du PMSI portant sur une autre pathologie.
- Etude de modèles de survie pour prendre en compte le délai de réhospitalisation (forêt aléatoire de survie, Cox boost, Cox Lasso, Cox Ridge).

L'ensemble des analyses seront réalisées dans un espace sécurisé. La construction de la base sera réalisée en SAS ou en langage SQL. Les modèles d'apprentissage seront implémentés au choix en R ou en python.

Contact :

Corinne Emery, Responsable du département biostatistiques

corinne.emery@cemka.fr

Stéphane Bouee, Responsable du département Real World Evidence

stephane.bouee@cemka.fr