CREATION D'UN ALGORITHME AUTOMATIQUE POUR VERIFIER L'HYPOTHESE DE POSITIVITE

Contexte

Le laboratoire SPHERE (methodS for Patients-centered outcomes and HEalth REsearch, INSERM UMR 1246, Université de Nantes, Université de Tours) et la société IDBC (groupe A2COM) ont décidé de créer ensemble le Laboratoire Commun RISCA (Research in Informatics and Statistics for Cohort-based Analyses, www.labcom-risca.com) afin de développer Plug-Stat®, un logiciel sur-mesure pour l'analyse et la valorisation des données observationnelles provenant d'une cohorte ou d'un registre.

Ce type de données permet des études dîtes en « vie réelle » (*Real World Evidence*) afin d'identifier le véritable impact clinique des interventions au sein du système de santé [1]. L'absence de randomisation dans ces études fait que la comparabilité des groupes n'est plus garantie par le design. Pour éviter des biais de confusion, il faut alors prendre en compte les facteurs de confusion lors de l'analyse. Les méthodes dites d'inférence causale permettent de pallier ce problème et d'estimer un effet causal à partir de données observationnelles.

Ces méthodes reposent sur plusieurs hypothèses dont la positivité. Equivalente à la clause d'ambivalence dans les essais cliniques randomisés, la violation de l'hypothèse de positivité survient lorsque certains patients de l'échantillon n'ont aucune (ou une très faible) chance d'être exposés ou non-exposés [2]. Cole et Hernan [3] ont proposé une représentation graphique basée sur le score de propension [4] afin de repérer d'éventuelles violations de cette hypothèse.

Objectifs

Les objectifs de ce stage sont les suivants :

- Réaliser un état de l'art des méthodes permettant de vérifier l'hypothèse de positivité.
- Développer un algorithme permettant la détection d'un problème de positivité pouvant biaiser l'estimation causale de l'effet d'intérêt.
- Identifier et représenter les caractéristiques des sujets responsables de la potentielle violation de l'hypothèse de positivité.
- Appliquer l'algorithme sur la cohorte AtlanREA (patients en réanimation médicale).

Description de la base de données

La base de données utilisée proviendra de la cohorte AtlanREA. Il s'agit d'une cohorte observationnelle prospective multicentrique française constituée de patients hospitalisés dans un service de soins intensifs. Deux populations principales sont présentes dans la cohorte : les patients traumatisés graves et les patients cérébrolésés. Dans chaque centre du réseau et pour chaque patient les caractéristiques à l'admission, mais aussi, les procédures, traitements, scores de gravité, événements indésirables sont

saisis quotidiennement jusqu'à la sortie de réanimation. Le devenir à la sortie de l'hôpital est également collecté.

Méthodologie envisagée

L'algorithme reste à définir, notre hypothèse de travail est d'utiliser un superlearner (SL) [5] afin de modéliser le score de propension. Le SL combine plusieurs méthodes d'apprentissage supervisé (ex : réseaux de neurones, forêts aléatoires, gradient boosting, etc.) en une seule. Pour obtenir une estimation finale, le SL pondère les estimations obtenues par les différentes méthodes [6]. Il est possible qu'il soit préférable de réduire le set de variables à introduire dans le SL aux variables associées à l'événement. Cette possibilité devra être étudiée, une pénalisation de type LASSO [7] pourrait alors être utilisée. Des simulations seront réalisées pour identifier et comparer les performances des différentes versions de l'algorithme qui auront pu être étudiées.

Une application de l'algorithme nouvellement développé sur la base de données AtlanREA illustrera son intérêt.

Résultats attendus

- Une revue de la littérature complète autour des méthodes permettant de représenter et diagnostiquer le non-respect de l'hypothèse de positivité.
- Un algorithme permettant la détection d'un problème de positivité.
- La rédaction d'une première version d'un article incluant une partie méthodologique avec la présentation de l'algorithme, l'étude de simulation et l'application en réanimation.

Profil attendu

Le candidat devra être étudiant en Master 2 Biostatistique ou équivalent. Il devra avoir une bonne connaissance des modèles linéaires généralisés. Il devra être à l'aise avec le langage de programmation R et avoir une forte appétence pour la programmation. Des connaissances en inférence causale et en machine learning seraient appréciées.

Structure d'accueil

L'étudiant sera accueilli dans le cadre du Laboratoire Commun RISCA (Research in Informatics and Statistics for Cohort-based Analyses, www.labcom-risca.com) qui est un partenariat entre l'unité INSERM U1246 SPHERE et l'entreprise IDBC. L'unité INSERM UMR 1246 – SPHERE (methodS for Patient-centered outcomes and HEalth REsearch) est une équipe de recherche multidisciplinaire dont l'objectif est de promouvoir la recherche méthodologique centrée sur le patient. IDBC, société du groupe A2COM, est une entreprise de services du numérique (ESN) spécialisée dans la conception

d'applications informatiques dans le secteur de la recherche médicale. Conjointement, les partenaires développent le logiciel d'analyse de données de santé Plug-Stat.

Les encadrants du stage sont localisés à l'IRS 2 à Nantes.

Lieu de stage : Institut de Recherche en Santé 2 (IRS2) – Université de Nantes – 22, Boulevard Bénoni-Goullin, 44200 Nantes

Encadrant principal : Arthur Chatton, Doctorant en biostatistique, IDBC.

Co-encadrant : Yohann Foucher, MCU en biostatistique, Université de Nantes.

Durée souhaitée : 6 mois

Période : de mars/avril à août/septembre 2021

Merci d'adresser votre CV et lettre de motivation à

Arthur Chatton
achatton@idbc.fr

Références

1. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? New England Journal of Medicine. 2016;375:2293–7.

2. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21:31–54.

3. Cole SR, Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. American Journal of Epidemiology. 2008;168:656–64.

4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

5. van der Laan M, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology. 2007;6:Article 25.

6. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. European Journal of Epidemiology. 2018;33:459–64.

7. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58:267–88.