**Suivi temporel de clusters de patients pour extraire les informations à partir des bases de données médico-administratives**

Judith Lambert1,2, Anne-Louise Leutenegger3, Anne-Sophie Jannot1,4, Anaïs Baudot2,5,6

*1 INSERM, Université Paris Cité, CRC, HeKA team, UMR1138, Paris, France*

*2Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France*

*3INSERM, Université Paris Cité, NeuroDiderot , UMR1141, Paris, France*

*4Department of Statistics, Medical Informatic and Public Health, HEGP, AP-HP*

*5CNRS, Marseille, France*

*6Barcelona Supercomputing Center, Barcelone, Espagne*

*Corresponding author: Judith Lambert, PariSanté Campus, 10 rue d'Oradour-sur-Glane, 75015 Paris, France, judith.lambert@inserm.fr*

**Introduction**

L’identification de clusters (sous-groupes) de patients à partir des bases de données médico-administratives est particulièrement importante pour mieux comprendre l’hétérogénéité des maladies. Cependant, ces bases de données contiennent différents types de variables longitudinales qui sont mesurées à différentes périodes, générant ainsi des données tronquées. Il est donc fondamental de développer des approches de clustering qui prennent en compte ces types de données. Dans notre étude, nous avons ainsi développé deux nouvelles approches permettant d’identifier des clusters de patients à partir des données longitudinales tronquées contenues dans les bases de données médico-administratives.

**Matériel et méthodes**

Nos deux approches sont basées sur une stratégie appelée cluster tracking où les clusters sont d’abord identifiés à chaque pas de temps et sont ensuite suivis au cours du temps. La première approche identifie les clusters directement à partir des données brutes alors que la deuxième commence par construire des réseaux de patients à partir des données brutes. Nous avons utilisé ces approches pour analyser les médicaments antithrombotiques prescrits entre 2008 et 2018 contenues dans l’Échantillon Généraliste des Bénéficiaires (EGB), une cohorte nationale française. Nous avons également comparé ces approches avec trois méthodes classiques de clustering longitudinal.

**Résultats**

Nos deux approches nous ont permis d’identifier des trajectoires de clusters correspondant à des pathologies. Elles ont également obtenu de meilleures performances comparé aux trois méthodes classiques de clustering longitudinal.

**Conclusion**

Les approches basées sur le cluster tracking permettent de prendre en compte les données longitudinales tronquées contenues dans les bases de données médico-administrative sans nécessiter d’imputation.