Quelle est la méthode la plus robuste face aux données manquantes ?

Vincent Leclerc*[1], Claire Le Corvaisier*[2], Laurent Bourguignon*[2], and Michel Ducher*[1]

[1]EMR3738 Ciblage thérapeutique en oncologie – Université Claude Bernard - Lyon I – France
[2]UMR5558 Equipe Evaluation et Modélisation des Effets Thérapeutiques – Université Claude Bernard - Lyon I (UCBL) – France

Résumé

Introduction : Modéliser le vivant, c'est faire face aux données manquantes. En effet, nous y sommes souvent confrontés en biologie et en santé. Cette étude propose une méthode permettant d'apprécier la robustesse de différents outils de modélisation face aux données manquantes.
Méthode : A partir d'une cohorte de 105 patients pédiatriques ayant reçu une greffe de moelle, nous avons créé une population virtuelle de 1000 patients dans laquelle nous avons généré des données manquantes (1%, 3%, 5%, 7%, 10%, 15%, et 20%). La robustesse de 5 types de modèles différents face à ces données manquantes a été comparée.

Résultats : Les réseaux bayésiens (RB) naïfs augmentés sont le modèle le plus robuste parmi les 5 testés. Ils conservent de très bonnes caractéristiques jusqu'à 10% de valeurs manquantes dans la base (AUC-ROC 0.79) contrairement à la régression logistique (AUC-ROC 0.73), au RB naïf (AUC-ROC 0.71), à la forêt aléatoire (AUC-ROC 0.72), et aux vecteurs machine (AUC-ROC 0.57).
Discussion – Conclusion : Nos résultats montrent une forte disparité de la robustesse des méthodes testés face aux données manquantes. Une réflexion initiale sur la robustesse de la méthode envisagée pour modéliser le vivant semble impérative, en particulier dans les secteurs où les données manquantes sont fréquentes.

Mots-Clés: Modélisation, Base de données, Clinique, Valeurs manquantes, Intelligence artificielle

*Intervenant