Gestion des données manquantes dans les grandes bases de données multicentriques de déformations rachidiennes : validation de la méthode MissForest.

Daniel Larrieu* [1], Cécile Roscop [1], Louis Boissière [2], Study Group European Spine [3], Yann Philippe Charles [4], Anouar Bourghli [5], Ibrahim Obeid [2], Alice Baroncini [6]

[1] Clinique Terrefort, Chirurgie de la déformation rachidienne, Bruges, France
[2] Clinique du Dos Bordeaux Terrefort, Orthopedics / Spine Surgery, Bruges, France
[3] Vall d'Hebron Research Institute, Spine Surgery Unit, Barcelona, Espagne
[4] CHRU Strasbourg, Chirurgie du Rachis, Strasbourg, France
[5] King Faisal Specialist Hospital and Research Center, Spine Surgery Department, Riyadh, Arabie Saoudite
[6] Humanitas San Pio X, Spine Surgery, Milan, Italie

INTRODUCTION

Les bases de données multicentriques en chirurgie du rachis comportent fréquemment une proportion importante de données manquantes, pouvant réduire la puissance statistique, introduire des biais et limiter les analyses longitudinales. Dans la base étudiée, le taux de données manquantes dépassait 60% à cinq ans de suivi, limitant la majorité des analyses au suivi à deux ans. L'objectif de cette étude était d'évaluer différentes méthodes d'imputation afin d'améliorer l'exploitation de données cliniques incomplètes issues d'un registre multicentrique de déformations rachidiennes.

MATÉRIEL ET MÉTHODE :

Nous avons analysé 1 633 patients avec un suivi à deux ans. Les variables présentant plus de 40% de données manquantes à plusieurs temps de suivi ont été exclues, laissant 112 variables analysables (104 quantitatives et 8 qualitatives). Les mécanismes de données manquantes ont été étudiés par analyses descriptives et tests statistiques. Un jeu de validation de 315 dossiers complets a été constitué, puis des valeurs manquantes artificielles ont été introduites afin de reproduire le schéma de données manquantes initial. Plusieurs méthodes d'imputation ont été comparées : imputation simple, kNN, MICE et MissForest. Les performances ont été évaluées par les indicateurs RMSE, MAE et MAPE. La méthode la plus performante a ensuite été appliquée à l'ensemble de la base.

RÉSULTATS :

Le taux global de données manquantes était de 19.7%. Lors de la validation, la méthode MissForest présentait les meilleures performances avec les plus faibles valeurs de RMSE, MAE et MAPE. Les distributions ont été préservées pour 96 des 104 variables quantitatives. De légères variations ont été observées pour 8 variables, et une variable (densité d'implants) présentait un effet de limitation d'intervalle. Appliquée à l'ensemble de la base comprenant 35 940 valeurs manquantes, la méthode MissForest conservait les moyennes, écarts-types et distributions globales sans biais systématique notable.

CONCLUSION :

Dans cette large base multicentrique de déformations rachidiennes, la méthode MissForest apparaît comme la stratégie d'imputation la plus performante, préservant la distribution des données dans 96 des 104 variables analysées. Cette approche constitue un outil robuste pour améliorer l'exploitation des bases longitudinales incomplètes et renforcer la qualité méthodologique des études cliniques en chirurgie du rachis.