

PRÉDICTION DYNAMIQUE INDIVIDUELLE D'ÉVÈNEMENT DE SANTÉ À PARTIR DE MULTIPLES DONNÉES LONGITUDINALES

Anthony Devaux ¹, Robin Genuer ^{1,2} & Cécile Proust-Lima ¹

¹ *INSERM U1219, Bordeaux Population Health, Bordeaux, FRANCE*

² *INRIA Bordeaux Sud-Ouest, Talence, FRANCE*

Introduction Les données individuelles collectées tout au long du suivi de patients sont des informations essentielles pour évaluer la probabilité de survenue d'un évènement de santé, et in fine adapter par exemple une stratégie thérapeutique. Des méthodes statistiques, basées sur la modélisation conjointe ou l'approche landmark, ont été proposées pour incorporer les données répétées d'un ou deux marqueurs dans des outils de prédiction dynamique, qui peuvent ensuite être mis à jour à chaque nouvelle information disponible. Ces méthodologies ne sont néanmoins pas adaptées pour un grand nombre de marqueurs. Dans ce travail, nous proposons donc une méthodologie pour la prédiction dynamique individuelle prenant possiblement en compte un très grand nombre de marqueurs répétés dans le temps. Nous combinons pour cela l'approche de landmark dynamique aux techniques d'apprentissage statistique adaptées aux données de survie.

Méthodes Le modèle landmark dynamique se décline en deux étapes. La première étape consiste à prédire des résumés individuels pour capter l'histoire de l'information de chacun des marqueurs au temps landmark choisi. Pour cela, nous utilisons les modèles mixtes uniquement sur les données collectées avant le temps landmark. Dans la deuxième étape, nous utilisons les résumés, obtenus dans la première étape, comme variables explicatives dans des modèles de survie adaptés à la grande dimension et aux variables corrélées (cox pénalisé, sparse-Partial Least Square, forêts aléatoires) uniquement sur les patients encore à risque après le temps landmark.

Applications Notre méthodologie a été appliquée à deux jeux de données. L'un pour prédire la survenue du décès chez les patients atteints de cirrhose primaire biliaire (maladie chronique du foie), et l'autre pour prédire la survenue du décès parmi des personnes âgées.

Simulations Nous avons simulé les données selon plusieurs scénarios en fonction du nombre de résumés associés à l'évènement ainsi que la forme de la fonction de dépendance (linéaire ou non linéaire) entre les variables explicatives et l'évènement. Dans les scénarios où l'association est non linéaire, les résultats ont montré que les forêts aléatoires en survie surpassent les autres méthodes en termes de Brier Score et AUC.