Prise en compte des groupes de biomarqueurs ou des voies biologiques dans les modèles de Cox pénalisés de haute dimension

Shaima Belhechmi[1, 2, *], Riccardo De Bin[3], Stefan Michiels[1, 2] & Federico Rotolo[4]

[1]Service de Biostatistique et d'Epidémiologie, Institut Gustave Roussy, F-94805, 114 Rue Edouard Vaillant, 94800 Villejuif, France.

[2]Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM U1018 Oncostat, F-94805, Villejuif, France.

[3]Department of Mathematics, University of Oslo, Oslo, Norway.

[4]Innate pharma, Marseille, France.

*Auteur correspondant : Shayma.BEL-HECHMI@gustaveroussy.fr (Shaima Belhechmi)

Résumé

Problématique Le développement de technologies génomiques à haut débit a permis la croissance rapide et la disponibilité plus facile de très grandes données génomiques. Le modèle à risques proportionnels de Cox est couramment utilisé pour estimer l'effet d'un ou de plusieurs facteurs pronostiques pour des critères de jugement de type survie. La méthode de régression pénalisée LASSO est utilisée pour sélectionner des biomarqueurs dans des données de haute dimension, mais cette méthode ne prend pas en compte les connaissances des rôles biologiques des biomarqueurs, par exemple les voies biologiques (pathways) connues.

Méthodologie Nous présentons différentes pénalisations pour que le modèle de Cox sélectionne correctement les biomarqueurs groupés afin de favoriser la sélection de biomarqueurs pronostiques (actifs) qui, en plus d'avoir un effet individuel important, appartiennent à un groupe actif. Nous avons considéré le cas des groupes pré-spécifiés et disjoints. Nous proposons la méthode Lasso Adaptatif avec des différents poids spécifiques pour chaque voie biologique. Nous avons comparé notre méthode proposée avec deux autres méthodes, le Sparse Group Lasso (SGL) et le Lasso Intégratif avec des facteurs de Pénalisation (IPF-Lasso). Pour l'approche Lasso Adaptatif, nous avons considéré six stratégies de pondération.

Nous avons évalué dans une étude de simulation la capacité de sélection (le taux de fausse découverte (FDR) et faux négatifs (FNR) ainsi que le FDR dans les groupes inactifs vs actifs) et de prédiction (l'air sous la courbe de ROC (AUC)) de ces méthodes. Nous avons illustré ces méthodes en utilisant des données d'expression de 109 gènes appartenant à trois voies (Système Immunitaire (47 gènes), Prolifération (43 gènes) et Stroma (19 gènes)) et de 614 patientes atteintes d'un cancer du sein traitées par chimiothérapie adjuvante.

Résultats Dans l'étude de simulation, les méthodes IPF-Lasso, SGL et Lasso Adaptatif avec la pondération du maximum de la statistique de Wald (MSW) présentaient la meilleure balance globale FDR-FNR. Les méthodes IPF-Lasso et SGL avaient le FDR le plus élevé dans les groupes inactifs (c'est-à-dire qui ne contient pas de biomarqueurs actifs).

Les propriétés favorables qui distinguent la méthode Lasso Adaptatif avec la pondération du MSW des autres méthodes sont le FDR le plus bas dans les groupes inactifs (entre 0.01 et 0.55 selon les scénarios) et l'AUC le plus grand par rapport à ces concurrents (entre 62% et 80%).

Dans l'application du cancer du sein, la méthode IPF-Lasso a sélectionné 14 gènes, dont un appartenant à la voie Stroma, deux à la voie Prolifération et 11 à la voie Immunitaire. La méthode SGL a sélectionné le modèle nul. La méthode Lasso Adaptatif avec la pondération MSW a sélectionné 3 gènes appartenant à la voie Prolifération.

Conclusion Nous préconisons la méthode Lasso Adaptatif avec la pondération du maximum de la statistique de Wald dans un modèle de Cox pénalisé.

Mots-clés Médecine stratifiée, données de grande dimension, régression pénalisée, biomarqueurs pronostiques, données génomiques, voies biologiques.

Déclaration d'intérêt Les auteurs déclarent n'avoir aucun conflit d'intérêts.