Sujet de thèse: Algorithmes statistiques pour la détection d'effets secondaires de cocktails de médicaments.

Etienne Birmelé, IRMA

Contexte

Les effets indésirables liés aux médicaments (ADR pour Adverse Drug Reactions) causent des millions d'hospitalisations évitables et des dizaines de milliers de morts par an à travers le monde. La plupart des ADR sont dues à combinaisons de médicaments, qui ne peuvent pas toutes être testées avant mise sur le marché.

Trouver des ADR après mise sur le marché et causés par des cocktails de médicaments nécessite de très grandes cohortes. Des calculs de puissance préliminaires montrent que des bases de plusieurs dizaines de millions de personnes sont nécessaires si deux médicaments sont pris par 5% de la population et qu'un ADR issu d'une interaction fait l'objet d'une remontée pour 1 patient sur 50 000. De plus, les médicaments peuvent être regroupés suivant une structure arborescente de plusieurs milliers de nœuds, les feuilles représentant les médicaments, et leurs parents de groupes de plus en plus larges (par exemple, l'hydrocortisone est un corticoïde oral local, les corticoïdes oraux locaux font partie des préparations stomatologiques, qui eux-mêmes font partie des médicaments à trait métabolique). La même structure arborescente est présente dans la classification des ADR (classification MEDDRA).

Le but du projet global est de faire progresser les méthodes de pharmacovigilances permettant de détecter des ADR et d'en faire bénéficier les agences du médicament. Pour cela, les étapes suivantes seront mises en place

- WP1 : Combiner les remontées d'ADR (effectuées auprès des agences du médicament par les médecins généralistes) avec la consommation de médicaments pour une base de données de plus de 100 millions de personnes issues de cinq pays européens (Norvège, Finlande, Allemagne, France, Italie) et constituer ainsi la plus grande base de données regroupant ordonnances et ADR. Mettre en place l'architecture qui permettra d'explorer efficacement ce jeu de données.
- WP2 : Développer des algorithmes de Machine learning permettant de détecter quels cocktails de médicaments favorisent quels types d'ADR. Pour cela, plusieurs tâches sont nécessaires :
 - 2.1. Définir la prise simultanée de médicaments. Une solution simple est de considérer leur présence sur la même ordonnance ou des ordonnances proches dans le temps, mais la réalité est plus complexe (ordonnances multiples, dosages, suivi du traitement)
 - 2.2 Définir de même la simultanéité entre prise de médicament et ADR.
 - 2.3. Définir un algorithme d'apprentissage statistique qui parcourt le plus intelligemment possible l'espace des cocktails afin de déceler les couples cocktails de médicaments (ou familles de médicaments)/ADR (ou famille d'ADR) qui sont significativement associés
- WP3 : Valider les approches du WP2 à l'aide de liaisons médicaments/ADR connues, plus ou moins fortes, et estimer ainsi la sensibilité et la spécificité de la méthode. Examiner et interpréter les nouvelles liaisons proposées d'un point de vue pharmaceutique.
- WP4 : Implémenter un outil dédié aux prescripteurs leur permettant, en fonction de l'historique de leur patient, d'évaluer les risques d'ADR d'une nouvelle prescription.

Ce projet général est porté par l’université arctique de Tromsø (Dr K. Svendsen, département de Pharmacie et L.A. Bongo, département d’informatique). Pour sa partie algorithmique et modélisation (WP2), il fait l’objet d’une collaboration avec l’université Paris Cité (V. Perduca, MAP5) et l’université de Strasbourg (E. Birmelé, IRMA).

Sujet

Le projet doctoral s’inscrit dans le cadre de la tâche WP2 du projet global décrit précédemment. Son objectif est l’élaboration d’un algorithme efficace (càd capable de tourner sur un grand jeu de données) permettant de proposer de nouveaux cocktails de médicaments ayant des effets indésirables, et que les pharmaciens pourront ensuite valider.

La première tâche consiste à explorer de manière efficace l’espace des possibilités, sachant qu’en plus de la très grande taille de l’espace des possibles, deux particularités sont à prendre en compte : 1) la structure arborescente en famille et sous-familles des médicaments et des ADR ; 2) l’absence d’observations pour de nombreux cocktails possibles, ce qui peut donner lieu à un ensemble de possibilités non connexes. Les algorithmes de Monte-Carlo génétiques sont une piste intéressante permettant à la fois de mener plusieurs explorations de front et de leur faire échanger des informations pertinentes. Il faudra cependant les adapter et étudier leurs propriétés dans le cadre présent. Leur implémentation se fera en C ou C++, avec un interfaçage en R ou Python pour pouvoir être utilisé facilement par les partenaires pharmaciens. La première étape du travail consistera à livrer un tel algorithme dans un cadre simplifié d’un jeu de données où les patients sont réputés utiliser un cocktail si les médicaments leur ont été prescrit dans un laps de temps fixé et où ne sont pris en compte que la population restreinte ayant fait remonter un ADR. Dans ce cas, à cocktail et ADR fixé, le score revient simplement à un odd-ratio utilisé classiquement en épidémiologie et basé sur des comptages (ADR cible ou non, prise du cocktail ou non). Ce premier algorithme sera validé sur jeu de données simulées, puis sur jeu de données réelles avec des couples cocktails/ADR connus afin d’estimer sa pertinence pour une utilisation à des fins d’exploration.

La seconde partie de la thèse consistera à enrichir le score donné en incluant une modélisation plus fine à discuter avec les pharmaciens, et permettant de prendre en compte : 1) le jeu de données entier, càd également les patients n’ayant pas fait remonter d’ADR, mais dont les prescriptions permettront d’estimer la prévalence des différents cocktails dans la population générale 2) un score cocktail/ADR plus fin permettant de prendre en compte les dosages des médicaments et l’incertitude liée au suivi du traitement prescrit. 3) d’éventuelles méthodes de parcours alternatives. L’algorithme enrichi sera implémenté et validé de la même manière que précédemment, puis appliqué à large échelle.

Premières références bibliographiques :

Svendsen, K. H. Halvorsen, S. Vorren, H. Samdal, and B. Garcia, “Adverse drug reaction reporting: how can drug consumption information add to analyses using spontaneous reports?,” *Eur. J. Clin. Pharmacol.*, vol. 74, no. 4, pp. 497–504, Apr. 2018.

Liang, Faming, and Wing Hung Wong. “Evolutionary Monte Carlo: applications to C p model sampling and change point problem.” *Statistica sinica* (2000): 317-342.

Liquet, B., Bottolo, L., Campanella, G., Richardson, S., & Chadeau-Hyam, M. (2016). R2GUESS: a graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses. *Journal of statistical software*, 69(2).

Robert, C. P., Casella, G., & Casella, G. (1999). *Monte Carlo statistical methods* (Vol. 2). New York: Springer.