Algorithmes statistiques pour la détection d'interactions médicamenteuses

Jules Bangard [1] & Étienne Birmelé [1]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg et CNRS 7 rue René-Descartes, 67000 Strasbourg, France,* {***jbangard***, ***birmele***}*@unistra.fr*

Résumé. Une méthode statistique computationnelle est proposée pour la détection d'interactions médicamenteuses à risque, mettant en avant les enjeux de la surveillance post-commercialisation des médicaments. Cette étude met en œuvre un algorithme de Monte Carlo par Chaîne de Markov (MCMC) et un algorithme génétique pour identifier ces interactions à partir de données de pharmacovigilance. Une analyse des performances de l'algorithme MCMC, réalisée sur des données simulées, a montré des résultats très satisfaisants.

Mots-clés. Interactions Médicamenteuses, Algorithme MCMC, Algorithme Génétique, Optimisation Combinatoire, Pharmacovigilance

Abstract. A computational statistical method is proposed for detecting at-risk drug-drug interactions, highlighting the challenges of post-marketing drug surveillance. This study implements a Monte Carlo Markov Chain (MCMC) algorithm and a genetic algorithm to identify these interactions from pharmacovigilance data. An analysis of the MCMC algorithm's performance, conducted on simulated data, showed very satisfactory results.

Keywords. Drug-Drug-Interactions, MCMC Algorithm, Genetic Algorithm, Combinatorial Optimization, Computational Model, Pharmacovigilance

1 Introduction

1.1 Problème

1.1.1 Pharmacovigilance

Les phases d'essais cliniques, cruciales pour l'autorisation de mise sur le marché des médicaments, sont souvent limitées en taille et en diversité de profils médicaux (enfants, personnes immunodéprimées ...). Malgré leur importance, ces essais peuvent ne pas révéler certains effets secondaires qui se manifestent seulement après une utilisation prolongée. Cette lacune souligne la nécessité d'un deuxième rempart, la pharmacovigilance, qui surveille les risques d'effets indésirables post-commercialisation (Wikipédia (2024)).

La pharmacovigilance repose sur la collecte d'informations provenant des professionnels de santé et des industriels à travers des rapports individualisés appelés ICSR (**I**ndividual

Case **S**afety **R**eport). Ces rapports contiennent des détails essentiels sur le patient, tels que son âge, ainsi que des informations sur les médicaments consommés et les effets rencontrés.

Grâce à l'émergence des différentes bases de données répertoriant ces ISCR, des méthodes d'exploitation de ces rapports ont émergé (Bate & Evans (2009)). Ces méthodes ont pour ambition la détection de "signaux", aiguillant ainsi les chercheurs en pharmacologie dans leur recherche de médicaments à risque.

1.1.2 Interactions médicamenteuses

Dans le domaine de la pharmacovigilance, l'attention s'est historiquement concentrée plus particulièrement sur l'identification d'effets secondaires résultant de la prise d'un unique médicament. Cependant, l'enjeu des interactions médicamenteuses est devenu un domaine de recherche de plus en plus étendu. Avec l'augmentation constante du nombre de médicaments disponibles, il devient impossible pour les pharmacologues d'examiner toutes les combinaisons possibles de manière exhaustive. Cette complexité est rendue préoccupante par certaines découvertes indiquant que divers traitements poly-médicamenteux offrent de meilleurs taux de rétablissement comparés aux traitements utilisant un unique médicament (Walkup et al. (2008)). Ainsi, il est essentiel de développer des méthodes capables d'évaluer le risque d'effets secondaires résultant de telles combinaisons. Pour cette raison, la méthode présentée a pour principal objectif la détection d'interactions médicamenteuses provoquant des effets secondaires chez les patients.

1.2 Données

1.2.1 Arbre des médicaments

Les médicaments sont organisés en arbre selon le système de classification Anatomique, Thérapeutique et Chimique (ATC) disposant de 5 niveaux de hiérarchie différents. Le plus haut niveau de hiérarchie est l'organe anatomique sur lequel agit le médicament, et le plus petit niveau la substance chimique classée. L'arbre des médicaments comptabilise un total de noeuds avoisinant les 5800. Dans la suite, les feuilles sont assimilées à un médicaments tandis que les autres noeuds à une famille de médicaments.

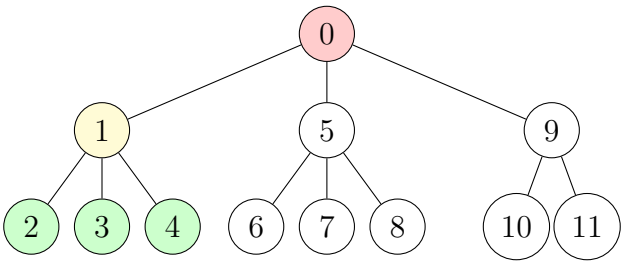

FIGURE 1 – Exemple simplifié d'arbre de médicaments

1.2.2 Cocktail de médicament

Pour représenter un cocktail de médicaments, l'arbre est numéroté d'après un algorithme de parcours en profondeur comme on peut le voir sur la Figure 1. Cette numérotation est utile par la suite pour encoder les cocktails. Un cocktail est une séquence de bits de taille n, où n est le nombre de noeuds dans l'arbre. Un bit à l'indice i vaut 1 si le médicament représenté par le noeud i de l'arbre est pris par le patient et 0 sinon. On a par exemple, pour un patient prenant les médicaments 2 et 8 de l'arbre 1, la séquence S suivante :

$$S = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Les bases de données de pharmacovigilance contenant les ICSR regroupent donc plusieurs séquences, toutes de taille n. Elles décrivent la prise de médicaments de chaque patient enregistré. Ces bases de données contiennent également les effets secondaires déclarés par les patients.

1.3 Caractérisation du risque associé à un cocktail

Ces données permettent d'associer un risque à une combinaison de médicaments. Il est possible de caractériser un tel risque à l'aide de différents scores. L'un des plus rependus de par sa facilité d'interprétation est le *Proportionnal Reporting Ratio* (PRR) proposé par (Evans et al. (2001)) et défini comme suit :

$$\frac{\mathbb{P}(AE|C)}{\mathbb{P}(AE|\neg C)}$$

où AE est un effet secondaire et C un cocktail de médicament.

Ce score est fréquemment utilisé en analyse de disproportionnalité (Bate et al. (1998); Evans, Waller, & Davis (2001); van Puijenbroek et al. (2002); Norén, Sundberg, Bate, & Edwards (2008)). Cette méthode utilise une table de contingence réalisée à l'aide d'une matrice regroupant les médicaments pris par chaque patient, ainsi qu'une matrice regroupant les effets secondaires subis par chaque patient encodée de la même manière que les cocktails de médicaments. Elle présente des avantages comme le temps de calcul qui est moindre mais également des inconvénients. Parmi eux, on peut citer les problèmes de masquages et de co-prescriptions (Maignen et al. (2014)).

De plus, certains scores ont été proposés pour généraliser ceux utilisés sur un unique cocktail comme le CRR ou le CSS qui sont des généralisations du PRR aux cocktails de taille supérieure à 1 (Noguchi et al. (2020)).

Une autre manière de caractériser le risque d'un cocktail est la suivante. Soient n le nombre de patient prenant le cocktail C, p la proportion de patient ayant l'effet secondaire AE et N le nombre total de patient dans le jeu de données. On définit X comme la nombre de personnes ayant pris le cocktail C et ayant subi l'effet secondaire AE. On prend comme score reflétant le risque

$$-\log(\mathbb{P}(X \geq x))$$

avec $X \sim \mathcal{H}(n, p, N)$ où $\mathcal{H}$ désigne la loi hypergeométrique.

2 Méthodes

L'identification de combinaisons de médicaments à haut risque est abordée via deux approches computationnelles. La première repose sur l'utilisation d'un algorithme de Monte-Carlo par chaînes de Markov (MCMC) pour l'exploration de l'espace des combinaisons de médicaments de taille p. Cette méthode permet d'estimer la distribution du risque associé à ces combinaisons. Ainsi il est possible de proposer une p-valeur empirique liée au score observé d'un cocktail. La seconde méthode, basée sur un algorithme génétique (Goldberg (2013)), vise à identifier de manière ciblée les combinaisons présentant un risque élevé, sans nécessiter une couverture exhaustive de l'espace des solutions.

2.1 Approximation du risque à travers les cocktails de médicaments

L'algorithme MCMC utilisé est l'algorithme de Metropolis-Hastings. Pour utiliser un tel algorithme, il faut définir un espace d'états $S = \{S_1, \ldots, S_p\}$. Il nécessite une mesure cible $f(S_i)$ calculable et des lois conditionnelles $q(.|S_i)$ sous lesquelles on sait simuler et grâce auxquelles il va pouvoir proposer de nouveaux états.

Ensemble d'états

Un état est décrit par un cocktail de médicaments comportant k médicaments, k étant un hyperparamètre qui n'évolue pas au cours de l'algorithme.

Les états explorés peuvent contenir des noeuds internes à l'arbre (représentant donc des familles de médicaments). Cela permet la détection de signaux plus généraux. Par exemple, le paracétamol pourrait renvoyer un faible signal tandis que si on remonte dans l'arbre, les antalgiques pourraient peut-être représenter un signal plus général. Ainsi tous les patients prenant au moins un médicament de cette famille de médicaments seront considérés.

Loi de proposition

Pour passer d'un cocktail à un autre, deux "mutations" différentes peuvent être effectuées, la mutation de type 1 et la mutation de type 2. Ces deux mutations sont complémentaires et exploitent la structure d'arbre des médicaments. Elles fonctionnent de la manière suivante.

Mutation de type 1 La mutation de type 1 consiste en un mouvement totalement aléatoire dans l'espace des cocktails.

Mutation de type 2 La mutation de type 2 consiste en un mouvement dit "local" vis-à-vis de la structure d'arbre des médicaments. En effet, lors de cette mutation on change un noeud de la séquence S_i en l'un de ses noeuds voisins libres.

La Figure 2 représente un exemple d'une mutation de type 1 et 2. Pour la mutation de type 2, initialement la séquence contient les noeuds 2 et 3 en vert. On aperçoit, en orange à l'aide des arêtes orientées, les mouvements que la séquence peut effectuer. Un mouvement (une arête) est tiré uniformément parmi ceux disponibles. Dans notre

exemple, l'arête allant du noeud 2 vers la feuille 6 est choisie, ainsi, le noeud 2 est supprimé de la séquence et le 6 quant à lui est ajouté.

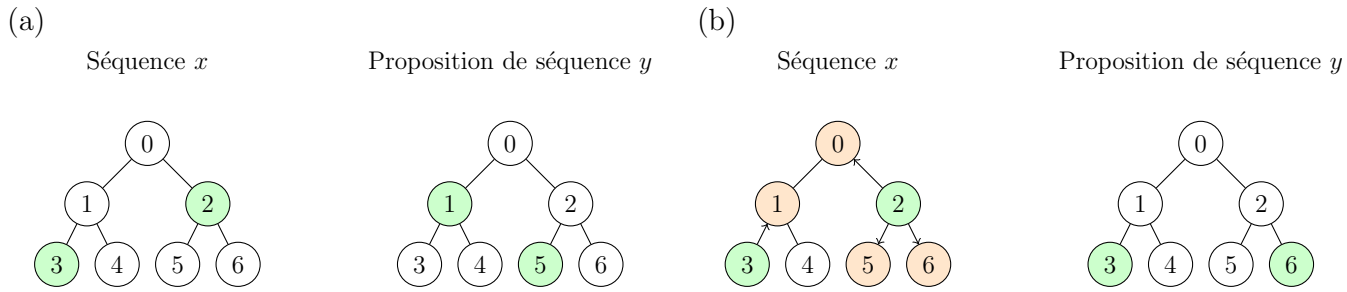

FIGURE 2 – (a) Exemple d'une mutation de type 1 (b) Exemple d'une mutation de type 2

La mutation de type 1 est proposée avec probabilité p_I à chaque itération, p_I étant un hyperparamètre. La mutation de type 2 est ainsi effectuée avec probabilité $p_{II} = 1 - p_I$.

Évaluation de l'état

L'évaluation d'un cocktail de médicaments repose sur l'un des scores présenté précédemment, noté $H(S)$. La mesure cible f choisie est alors la suivante :

$$f_T(S_i) = \frac{1}{Z(T)} \times e^{\frac{H(S_i)}{T}}$$

Où $Z(T) = \sum_S e^{\frac{H(S)}{T}}$. T est un paramètre appelé température qui permet de moduler l'exploration de l'espace en acceptant plus aisément les cocktails de score modéré (T élevé) ou, au contraire, en privilégiant fortement les combinaisons de médicaments de score élevé (T faible).

La probabilité d'acceptation du cocktail S_{i+1} à partir du cocktail S_i est :

$$\min(1, \frac{f_T(S_{i+1})}{f_T(S_i)} \times \frac{q(S_i|S_{i+1})}{q(S_{i+1}|S_i)})$$

La théorie liée à l'algorithme de Metropolis-Hastings assure que la loi empirique des $f(S_i)$ pour la chaîne de cocktail ainsi construite converge vers la loi de $f(S)$. Une réalisation très longue d'une telle marche permet donc d'obtenir une loi approchée qui permet de déterminer une p-valeur empirique pour le score d'un cocktail d'intérêt.

2.2 Recherche des cocktails présentant le plus gros risque

L'algorithme génétique suit le modèle habituel de ce type d'algorithmes comme le montre la Figure 3. Il fait évoluer une population dans le but d'obtenir comme résultat une population performante au vu d'un critère d'évaluation arbitraire. Les étapes nécessaires pour cela sont :

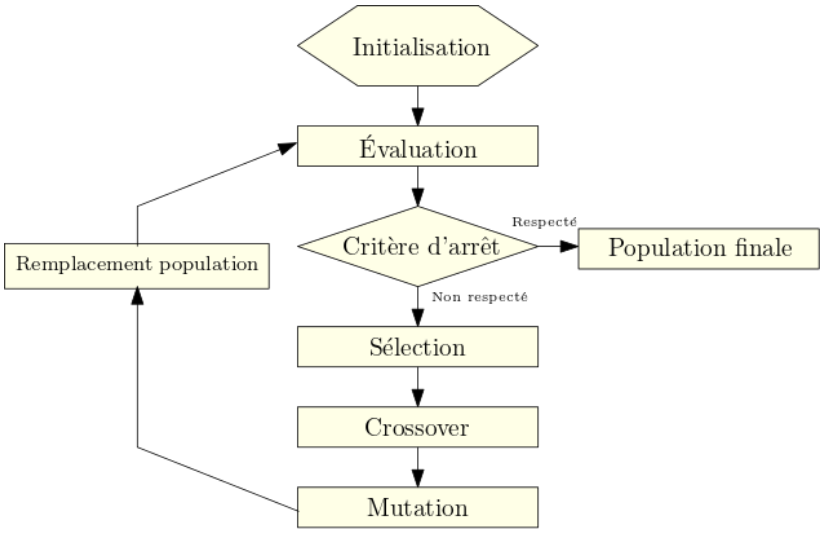

FIGURE 3 – Déroulement de l'algorithme génétique

Initialisation La population de l'algorithme génétique est un ensemble de m cocktails de médicaments. Ces cocktails sont initialisés de manière aléatoire et peuvent être de tailles différentes contrairement à l'algorithme MCMC.

Évaluation & Sélection À chaque itération, la population passe à travers une phase d'évaluation et de sélection. L'évaluation calcule, pour chaque cocktail, son score hypergéométrique présenté en partie 1.3. Les scores des cocktails qui se ressemblent au sein de la population sont ensuite pénalisés. Le but de cette pénalisation est d'obtenir une population qui n'est pas trop homogène de sorte à obtenir plusieurs combinaisons médicamenteuses préoccupantes en sortie.

Les scores ainsi obtenus permettent d'effectuer des tournois consistant à tirer k individus de la population et à conserver le meilleur des k pour la phase de reproduction. De tels tournois sont effectués jusqu'à obtenir le nombre d'individus désiré pour la phase de reproduction.

Modification & Remplacement de population L'évolution de la population vers une population performante vis à vis du critère d'évaluation se fait en deux temps.

Dans un premier temps, une opération appelée crossover permet à deux séquences d'échanger de l'information. Dans le cas présent, le crossover consiste en l'échange de sous-arbres entre deux séquences de la manière suivante :

— Un noeud **interne** v de l'arbre est sélectionné aléatoirement.

— Les noeuds du sous arbre de racine v sont échangés entre les deux séquences.

Après avoir effectué ce crossover, une mutation est appliquée aux individus résultants, choisie parmi deux possibilités. La première est la mutation de type 2 vue dans la section 2.1. La seconde fonctionne de la manière suivante, en notant p la longueur de la séquence et α un hyperparamètre à choisir :

— Avec probabilité $\frac{\alpha}{p}$ un noeud de l'arbre tiré uniformément est ajouté à la séquence .
— Avec probabilité $1 - \frac{\alpha}{p}$ un médicament de la séquence, tiré uniformément, est retiré.

Critère d'arrêt L'algorithme prend fin lorsque le critère est respecté. Dans notre cas, il s'agit d'un nombre d'itérations fixé par l'utilisateur.

3 Application et premiers résultats

3.1 Jeu de données simulé

Plusieurs jeux de données simulés ont été générés pour évaluer la performance de l'algorithme de Metropolis-Hastings. Pour ces jeux de données, plusieurs réponses sont définies préalablement, correspondant à divers cocktails de médicaments à risque. Chaque combinaison de médicaments à risque est d'intensité différente, c'est à dire, correspond à une probabilité plus ou moins faible de subir l'effet secondaire. Des informations ont étés recueillies auprès de pharmacologues dans le but de proposer un jeu de données simulé réaliste. Ce jeu de données comporte 200.000 patients. Il y a trois cocktails d'intérêt au sein de cette population. Chaque cocktail est pris par 1% de la population et provoque l'effet secondaire considéré avec une probabilité différente. Les trois probabilités sont $\frac{1}{50}$, $\frac{1}{100}$ et $\frac{1}{500}$. Un cocktail aléatoire donne l'effet secondaire d'intérêt avec probabilité $\frac{1}{2000}$.

3.2 Choix du score d'intérêt

Dans un premier temps. Une comparaison de la pertinence des différents indices de la section 1.3 est menée à l'aide de ces simulations.

Les différents indices ont été calculés sur 30 cocktails donnant lieu à un effet secondaire et 100.000 cocktails n'en donnant pas. Après classement par ordre croissant, cela permet de tracer des courbes Precision Recall présentées en Figure 4. Ces courbes sont identiques pour les trois méthodes RR, CRR et CSS, avec de très mauvais résultats. La méthode hypergéométrique donne de bien meilleurs résultats comme illustré également en Figure 5. En effet, les cocktails de plus haut score sont de vrais positifs, tandis qu'il s'agit de faux positifs pour les trois autres indices. Sur la Figure 5, les points de même abscisse sont légèrement décalés les uns par rapport aux autres sur l'axe des ordonnées de sorte à mieux visualiser les zones contenant beaucoup de cocktails.

L'analyse de disproportionnalité à travers les cocktails de médicaments est une tâche non triviale pour plusieurs raisons, l'une d'entre elles étant le bruit. Par bruit, on entend ici les ensembles de médicaments étant pris par peu de personnes dans le jeu de données. Certaines d'entre elles subissent l'effet secondaire étudié. Cela entraîne des valeurs de risques remarquablement élevées (voir CRR Figure 5), tandis qu'une conclusion sur la dangerosité du cocktail concerné semble impossible en raison de la taille de l'échantillon. De plus, le risque relatif attribue la même valeur à un cocktail consommé par 3 personnes, dont une

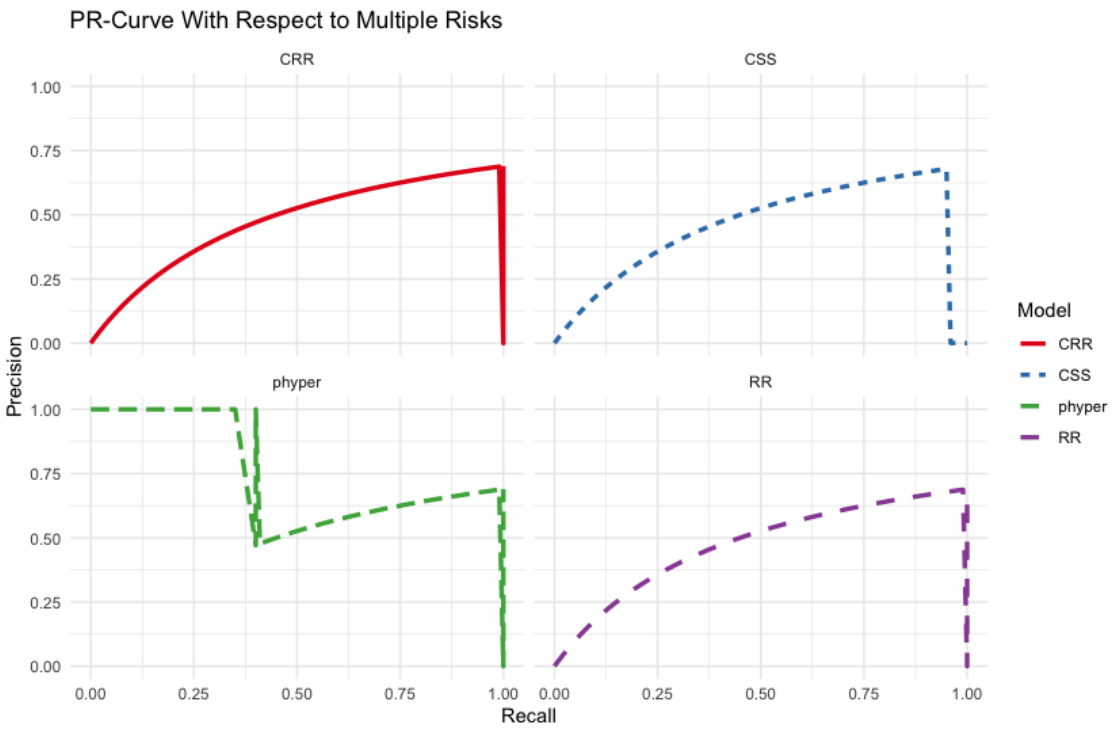

FIGURE 4 – Comparaison des courbes Precision-Recall associées aux différents risques pour un jeu de données simulé

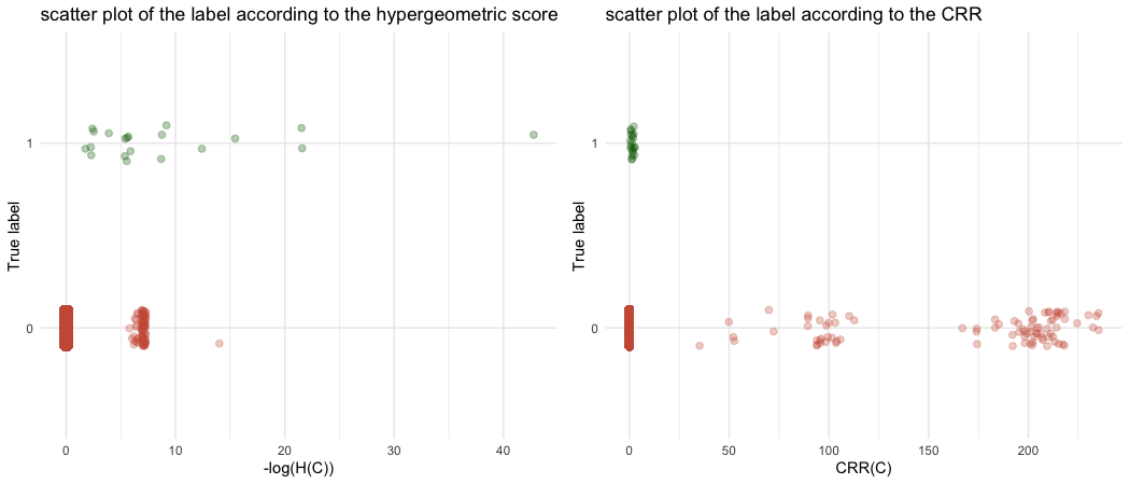

FIGURE 5 – Scatter plot montrant les scores des cocktails en fonction de leur vrai label, à gauche pour le score hypergéométrique, à droite pour le CRR

subit l'effet secondaire, qu'à un cocktail pris par 30 personnes, avec 10 d'entre elles subissant l'effet secondaire. Or, ce dernier est vraisemblablement plus à risque que le cocktail peu pris.

Le score hypergéométrique permet de prendre cet aspect en compte en attribuant un risque plus élevé aux combinaisons étant prises par un plus grand nombre de personne pour une même proportion d'effets secondaires observés.

3.3 Approximation de la distribution du risque à travers les cocktails de médicaments

L'algorithme MCMC permet d'obtenir une estimation de la distribution du risque à travers la population des cocktails d'une taille fixée. Dans notre exemple, les simulations sont effectuées sur des cocktails de taille 2 en utilisant le score hypergéométrique. Ainsi, il est possible de calculer la distribution réelle du risque en explorant de manière exhaustive toutes les combinaisons médicamenteuses de taille 2. Au delà, la combinatoire des cocktails devient trop grande pour calculer la distribution exhaustivement en un temps raisonnable.

La Figure 6 présente l'histogramme des deux distributions en ne gardant que les cocktails de risque non nul, ainsi que le QQ-plot associé. Les cocktails de risque nul sont éliminés car ils sont majoritaires au point d'écraser le reste de l'histogramme. L'approximation ainsi que la distribution réelle se trouvent respectivement en haut à gauche et en bas à gauche de la Figure 6. On constate que l'approximation en dimension deux est satisfaisante. On remarque de plus que la majorité de la distribution se situe aux alentours de 0.

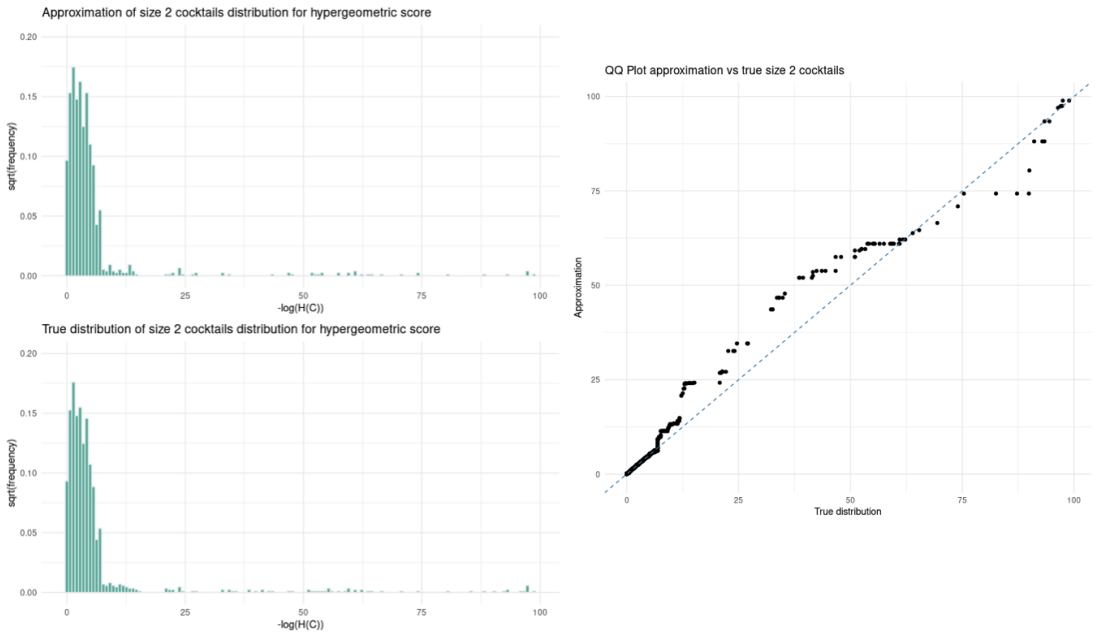

FIGURE 6 – Comparaison de la distribution réelle à la distribution estimée, conditionnellement à un risque non nul, à l'aide d'un diagramme quantile-quantile

Le QQ-plot n'est pas parfait mais il est à noter que 99% des scores correspondent à des points dans le segment initial en bas à gauche de la figure.

3.4 Algorithme génétique et données réelles

L'implémentation de l'algorithme génétique ainsi qu'une application des méthodes développées au jeu de données FAERS (Food & Administration (2024)), qui est un jeu de données publique de la Food & Drug Administration, sont en cours.

Bibliographie

Bate, A., & Evans, S. (2009). Quantitative signal detection using spontaneous adr reporting. *Pharmacoepidemiology and drug safety*, *18*(6), 427–436.

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A bayesian neural network method for adverse drug reaction signal generation. *European journal of clinical pharmacology*, *54*, 315–321.

Evans, S. J., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and drug safety*, *10*(6), 483–486.

Food, & Administration, D. (2024). *Fda adverse event reporting system (faers) quarterly data extract files.* Consulté sur `https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html` ([En ligne ; Page disponible le 31/01/2024])

Goldberg, D. E. (2013). *Genetic algorithms*. pearson education India.

Maignen, F., Hauben, M., Hung, E., Holle, L. V., & Dogne, J.-M. (2014). A conceptual approach to the masking effect of measures of disproportionality. *Pharmacoepidemiology and drug safety*, *23*(2), 208–217.

Noguchi, Y., Aoyama, K., Kubo, S., Tachi, T., & Teramachi, H. (2020). Improved detection criteria for detecting drug-drug interaction signals using the proportional reporting ratio. *Pharmaceuticals*, *14*(1), 4.

Norén, G. N., Sundberg, R., Bate, A., & Edwards, I. R. (2008). A statistical methodology for drug–drug interaction surveillance. *Statistics in medicine*, *27*(16), 3057–3070.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G., Lindquist, M., Orre, R., & Egberts, A. C. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety*, *11*(1), 3–10.

Walkup, J. T., Albano, A. M., Piacentini, J., Birmaher, B., Compton, S. N., Sherrill, J. T., ... others (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine*, *359*(26), 2753–2766.

Wikipédia. (2024). *Pharmacovigilance — wikipédia, l'encyclopédie libre.* Consulté sur `http://fr.wikipedia.org/w/index.php?title=Pharmacovigilanceoldid=204757946` ([En ligne ; Page disponible le 31/01/2024])