Projet de thèse

Intitulé de la thèse

Détection automatisée de signaux en pharmacovigilance : Exploitation conjointe de données de notifications spontanées et médico-administratives.

Spécialité du doctorat

Biostatistique, ED420 « Santé Publique »

Problématique scientifique générale

La pharmacovigilance a pour objectif de détecter le plus précocement les effets indésirables de médicaments commercialisés. Elle repose le plus souvent sur l'exploitation de notifications spontanées, c'est à dire la déclaration par un professionnel de santé de la survenue d'un événement indésirable dont l'origine suspectée est médicamenteuse.

A l'échelle nationale ou internationale, ces notifications spontanées représentent de grands ensembles de données. Il a donc été proposé depuis une quinzaine d'années un certain nombre de méthodes statistiques visant à détecter des associations statistiques suspectes entre médicaments et effets indésirables (Almenoff et al. 2007; Harpaz et al. 2012). De manière schématique, pour un couple médicament-événement (i,j) donné, les mesures d'associations proposées dans la littérature reposent généralement sur l'agrégation des données de notifications sous la forme de la table de contingence 2x2 croisant la présence/absence du médicament i avec la présence/absence de l'événement j.

Une mesure d'association est calculée simultanément pour chacun des couples présents dans la base. Un signal est déclenché lorsque la mesure d'association dépasse un certain seuil.

Ces approches sont donc caractérisées par le fait que les individus servant de référence sont des personnes ayant rencontré un autre événement indésirable et/ou pris un autre médicament que le couple (i,j) et ayant fait l'objet d'une notification spontanée par un professionnel de santé.

Malgré le caractère obligatoire de la notification spontanée, des travaux ont montré que la sous-notification est très importante même dans le cas d'effets indésirables graves. De plus, cette sous-notification varie en fonction du type de médicaments, du patient ainsi que du type d'événement indésirable. Ainsi, on parle de détection de signaux, les signaux statistiques générés à partir de ces données devant obligatoirement être évalués par des experts pharmacovigilants.

Plus récemment, un intérêt croissant s'est porté sur l'exploitation d'autres sources données pour la génération de signaux. En particulier, aux Etats Unis, un projet de très grande envergure s'est focalisé sur l'exploitation de grandes bases privées de données de remboursement de médicaments et dont les conclusions ont été publiées dans un supplément de la revue Drug Safety (Drug Safety, 2013, Vol. 36, Supplement 1 (pp. S1-S204). Studying the Science of Observational Research: Empirical Findings from the Observational Medical Outcomes Partnership).

En France, l'exploitation des données du SNIIRAM (Système national d'information inter-régimes de l'Assurance maladie) dans le cadre d'études pharmacoépidémiologiques a fait

l'objet de travaux récents (Weill et al. 2010). La richesse de cette base de données, l'une des plus grandes bases médico-administratives au monde, réside dans le fait qu'elle contienne l'ensemble des remboursements de soins pour la presque totalité de la population française. De plus ces données sont aussi chaînées au PMSI (Programme de médicalisation des systèmes d'information), c'est-à-dire aux séjours hospitaliers. L'accès à ces données est fortement réglementé et nécessite des demandes spécifiques à la CNIL. Néanmoins, la CNAMTS a fortement facilité l'accès de la communauté de chercheurs à l'échantillon généraliste des bénéficiaires (EGB) rassemblant les données du SNIIRAM pour un échantillon au 1/97^{ème} de la population française.

L'évaluation de l'intérêt de l'EGB pour la détection de signaux dans le cas d'effets indésirables graves (nécessitant une hospitalisation) fait l'objet d'un projet démarré au premier janvier 2015 au sein de l'équipe (Génération automatisée de signaux de pharmacovigilance sur l'échantillon généraliste de bénéficiaires avec la méthode en série de cas : développements méthodologiques et étude pilote, Financement ANSM, PI Ismaïl Ahmed). La difficulté dans l'exploitation de ces grandes bases de données médico-administratives réside dans le fait qu'elles n'ont pas été conçues avec un objectif de réponse à des questions de santé. Ainsi le SNIIRAM contient très peu de données sociodémographiques. Il ne contient aussi aucune information directe concernant les pathologies pour lesquelles sont dispensés les soins (en dehors des affections longues durées). Par ailleurs les effets indésirables pouvant être étudiés sont nécessairement « graves » puisque requérant une hospitalisation. Enfin, l'accès aux données du SNIIRAM dans un temps compatible avec de la détection de signaux est restreint à l'EGB. Malgré une taille très importante (environ 700000 personnes), son exploitation pour la détection d'effets indésirables n'est réalisable que pour des effets indésirables relativement fréquents dans la population.

Objectifs scientifiques de la thèse

L'objectif général de ce projet de thèse est de proposer de nouveaux outils de détection automatisée de signaux tirant partie des complémentarités des sources d'information que sont les notifications spontanées et l'EGB.

La première étape de ce travail de thèse consistera à développer une méthode de détection de signaux s'appuyant à la fois sur les données de notifications spontanées de la base nationale de pharmacovigilance française et les données de l'EGB. L'originalité de cette approche est qu'elle utilisera les données de l'EGB pour la constitution d'un groupe contrôle tandis que les cas seront extraits à partir des notifications spontanées. Dans un premier temps, nous envisagerons l'utilisation d'approches appariées « univariées » : pour un médicament étudié, les témoins seront appariés aux cas selon leur âge et leur sexe. Nous réfléchirons aussi à des appariements prenant en compte le profil d'expositions médicamenteuses autres que celui étudié. Cette méthode s'appuiera donc vraisemblablement sur un modèle de régression logistique conditionnelle. Le seuil de significativité des associations trouvées devra aussi tenir compte du nombre important de tests réalisés par l'utilisation de critères statistiques tels que le False Discovery Rate (Benjamini & Hochberg 1995). L'évaluation de cette nouvelle méthodologie de détection sera conduite sur la base de plusieurs ensembles existants de signaux de référence établis de différentes manières : expertise humaine (Ahmed et al. 2012; Ryan et al. 2013) exploitation de la littérature

biomédicale (Xu & Wang 2014), exploitation du résumé des caractéristiques du produit (Kuhn et al. 2010; Candore et al. 2015).

Dans un deuxième temps nous nous attacherons à comparer l'intérêt de cette méthodologie par rapport à une détection automatisée « classique », c'est-à-dire s'appuyant uniquement sur les données de notifications spontanées. La comparaison sera basée sur les différents ensembles de signaux de référence mentionnés plus haut, ainsi que sur une étude de simulations.

Le troisième temps de ce travail sera consacré à l'exploration de méthodes « multivariées » comme alternative aux méthodes appariées. Le nombre de médicaments mentionnés pour un événement indésirable donné étant potentiellement très grand, ce volet pourra faire appel à des approches telles que les régressions pénalisées de type lasso (Tibshirani 1996; Genkin et al. 2007) qui commencent à être explorées en pharmacovigilance (Caster et al. 2010). Néanmoins, ces méthodes ont été initialement développées avec un objectif prédictif. Leur utilisation dans un but de sélection de variables ; en pharmacovigilance, il s'agit d'identifier un nombre restreint de médicaments associés à un effet indésirable donné; fait l'objet de développements méthodologiques actuels (Meinshausen & Bühlmann 2010). Plus généralement, les méthodes issues de l'apprentissage statistique n'ont pas encore été beaucoup abordées en pharmacovigilance. Ce travail de thèse pourra donc être l'occasion d'explorer certaines d'entre elles. En particulier les Random Forest (Breiman 2001) qui ont récemment été revisités dans un contexte de sélection de variables semblent être une piste intéressante (Genuer et al. 2010; Hapfelmeier & Ulm 2013).

Titre en anglais du projet de thèse

Automated signal detection in pharmacovigilance from large spontaneous reporting and medico-administrative databases

Profil du candidat

Le candidat sera titulaire d'un diplôme d'école d'ingénieur en statistique /mathématiques appliquées (ENSAE, ENSAI, ISUP, INSA, Agro, Centrale, ENSIMAG,...) ou d'un M2 de statistique ou biostatistique. Sa formation comprendra un solide cursus théorique en statistique et un goût et des aptitudes pour les applications biomédicales et le travail collaboratif avec d'autres disciplines scientifiques, ainsi que de fortes compétences en programmation.

Laboratoire d'accueil

La thèse se déroulera dans l'équipe « Biostatistique et Pharmacoépidémiologie » de l'unité Inserm U1181 « Biostatistique, Biomathématique, Pharmacoépidémiologie et Maladies Infectieuses » (B2PHI), une unité de recherche mixte Inserm, UVSQ et Institut Pasteur.

L'unité 1181 est dirigée par Didier Guillemot et comprend 2 équipes. L'équipe Biostatistique et Pharmacoépidémiologie, localisée à Villejuif, est dirigée par Pascale Tubert-Bitter. Cette équipe a notamment développé de nombreux travaux statistiques pour la détection de signal en pharmacovigilance à partir des systèmes de notifications spontanées, ainsi qu'une expertise pour l'exploitation du SNIIRAM en pharmacoépidémiologie.

L'encadrement de la thèse sera conjoint entre Ismaïl Ahmed (Chargé de recherche Inserm) et Pascale Tubert-Bitter (Directeur de recherche Inserm).

Financement

Le financement sera sollicité auprès de l'école doctorale. Les modalités de candidature sont détaillées sur le site de l'ED420 (<http://www.ed-sante-publique.u-psud.fr/etre-candidat/trouver-un-financement/contrats-doctoraux-ed420.html>).

Contact

Ismail Ahmed : ismail.ahmed@inserm.fr

Pascale Tubert-Bitter : pascale.tubert@inserm.fr

Bibliographie

Ahmed, I. et al., 2012. Early detection of pharmacovigilance signals with automated methods based on false discovery rates: a comparative study. *Drug Saf*, 35(6), pp.495–506.

Almenoff, J.S. et al., 2007. Novel statistical tools for monitoring the safety of marketed drugs. *Clin Pharmacol Ther*, 82(2), pp.157–166.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 57, pp.289–300.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5–32.

Candore, G. et al., 2015. Comparison of Statistical Signal Detection Methods Within and Across Spontaneous Reporting Databases. *Drug Saf*.

Caster, O. et al., 2010. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. *Statistical Analysis And Data Mining*, 3(4), pp.197–208.

Genkin, A., Lewis, D.D. & Madigan, D., 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, 49(3), pp.291–304.

Genuer, R., Poggi, J.-M.-. M. & Tuleau-Malot, C., 2010. Variable selection using random forests. *Pattern Recognition Letters*, 31(14), pp.2225–2236.

Hapfelmeier, A. & Ulm, K., 2013. A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, 60, pp.50–69.

Harpaz, R. et al., 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther*, 91(6), pp.1010–1021.

Kuhn, M. et al., 2010. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6, p.343.

Meinshausen, N. & Bühlmann, P., 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), pp.417–473.

Ryan, P.B. et al., 2013. Defining a reference set to support methodological research in drug safety. *Drug Saf*, 36 Suppl 1, pp.S33–47.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 58, pp.267–288.

Weill, A. et al., 2010. Benfluorex and valvular heart disease: a cohort study of a million people with diabetes mellitus. *Pharmacoepidemiology and drug safety*, 19(12), pp.1256–62.

Xu, R. & Wang, Q., 2014. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *J Biomed Inform*.