Machine learning et big data en santé : le partenariat entre la Caisse Nationale d'Assurance Maladie et l'Ecole polytechnique

MOTS-CLÉS : MACHINE LEARNING, BIG DATA, PHARMACOVIGILANCE

Machine learning and big data in health : the partnership between Caisse Nationale d'Assurance Maladie and Ecole polytechnique

KEY-WORDS: MACHINE LEARNING, BIG DATA, PHARMACOVIGILANCE

Emmanuel BACRY*, Stéphane GAÏFFAS**

Le/les auteur(s) déclare(nt) n'avoir aucun lien d'intérêt en relation avec le contenu de cet article.

RÉSUMÉ

La Caisse Nationale d'Assurance Maladie (Cnam) et l'Ecole polytechnique ont signé fin 2014 une convention de partenariat de recherche pour une durée de 3 ans qui a été reconduite pour 3 ans début 2018. L'objectif de ce partenariat est de développer des technologies big data et d'apprentissage machine sur la base de données du Système National Inter-Régimes de l'Assurance Maladie (Sniiram). Nous décrivons dans cet article les principaux accomplissements de ces trois premières années de partenariat. Accomplissements technologiques d'une part : mise en place d'une pipeline big data de traitement des données et accomplissements méthodologiques d'autre part : développement de nouveaux algorithmes pour un dépistage automatisé de médicaments ayant des effets secondaires néfastes. La pertinence de ces outils a été confirmée par la détection "en aveugle" du pioglitazone, antidiabétique retiré du marché en 2011 pour cause d'augmentation du risque de cancer de la vessie. Nous concluons cet article en présentant les projets en cours et les perspectives de déploiement à plus grand échelle de ces algorithmes.

SUMMARY

The Caisse Nationale d'Assurance Maladie (Cnam) and the Ecole Polytechnique signed at the end of 2014 a research partnership agreement for a period of 3 years which was renewed for 3 years at the beginning of 2018. The objective of this partnership is to develop big data technologies and machine learning based on data from the Système National Inter-Régimes de l'Assurance Maladie (Sniiram). describe in this article the main achievements of these first three years of partnership. Technological achievements on the one hand: implementation of a big data processing pipeline and methodological achievements on the other hand: development of new algorithms for automated screening of drugs with adverse side effects. The relevance of these tools was confirmed by the "blind" detection of pioglitazone, an antidiabetic drug withdrawn from the market in 2011 due to its increasing of the risk of bladder cancer. We conclude this article by presenting ongoing projects and the prospects for the deployment of these algorithms on a larger scale

* CNRS, CEREMADE Université Paris-Dauphine PSL, 75775 Paris Cedex 16 et CMAP Ecole polytechnique, 91128 Palaiseau Cedex ,. E-mail : emmanuel.bacry@polytechnique.edu
** LPSM, Université Paris Diderot et CMAP Ecole polytechnique, LPSM, Université Paris bâtiment Sophie Germain, 5 rue Thomas Mann Diderot, 75205 Paris Cedex 13. Email : stephane.gaiffas@lpsm.paris

1. Introduction

Fin 2014, la Caisse nationale de l'assurance maladie (Cnam) et l'Ecole polytechnique ont signé une convention de partenariat de recherche et développement pour une durée de 3 ans. Reconduit jusqu'à la fin de l'année 2020, ce partenariat a pour objectif de favoriser le développement des technologies du big data appliqué au domaine de la santé. Plus précisément, cette collaboration a pour ambition de déployer de nouvelles pistes d'exploitations des données du Système National Inter-Régimes de l'Assurance Maladie (Sniiram). Cette base de données rassemble essentiellement les données de remboursements et d'hospitalisation des bénéficiaires de l'ensemble des régimes d'assurance maladie obligatoire en France (*cf Article de C. Gissot dans le même volume*). Il ne s'agit pas à d'une base de données cliniques mais d'une base de "dossiers de santé électroniques" (Electronic Health Records - EHR). Ces bases sont d'une grande richesse et leur analyse est devenue un sujet d'étude à part entière, source de très nombreux articles de recherche [2, 5, 7, 9]. La France, à travers le Sniiram, a la chance de posséder l'une des bases EHR les plus volumineuses au monde. Avec plus de 65 millions de dossiers de santé, 1,2 milliard de remboursements par an et 11 millions de séjours hospitaliers, cette base pèse plus de 200 To. D'une richesse inouïe, son analyse est de toute première importance tant pour les impacts potentiels en santé qu'en économie (le budget de la santé publique est le premier budget de l'état Français). Les équipes de statisticiens de la Cnam s'y attèlent depuis plusieurs années et ont obtenu de nombreux résultats de première importance à l'instar des résultats obtenus en 2013 sur le risque thromboembolique des pilules de 3ème génération.

L'infrastructure machine actuelle, l'organisation des données et les solutions logicielles de la Cnam ont été pensées pour optimiser l'objet initial du Sniiram, à savoir le remboursement des soins, mais sont peu adaptés à la recherche méthodologique. Cela constitue un facteur limitant fortement la fouille systématique des données à grande échelle, fouille devenue le standard du monde du big data et de l'intelligence artificielle, explorations de territoires encore inconnus et promesses d'innovations.

Ce partenariat s'est donné pour but le développement d'algorithmes définis au regard des missions de la Cnam et plus largement des enjeux de santé publique. Des algorithmes disruptifs pour la détection de signaux faibles ou anomalies en pharmaco-épidémiologie, pour l'identification de facteurs utiles à l'analyse des parcours de soins ou encore pour la lutte contre les abus et la fraude. De tels développements ont nécessité de repenser l'infrastructure de la Cnam, infrastructure machine, organisation des données et des outils d'analyse. Ce sont les premiers résultats des trois premières années de cette expérience tout à fait unique, tant

par l'ampleur de la tâche que par ses impacts sociétaux potentiels, que nous nous proposons de décrire succinctement ici. Il s'agit d'un travail très largement collectif, bénéficiant d'un cadre multidisciplinaire exceptionnel, rassemblant data-scientists, développeurs, chercheurs en mathématiques ou informatique et experts métiers du Sniiram ainsi que médecins spécialistes en santé publique.

2. L'infrastructure de la Cnam revisitée

Comme évoqué précédemment, l'infrastructure actuelle de la Cnam n'est pas pensée pour faciliter des analyses statistiques à grande échelle. Cela constitue un frein à la recherche méthodologique. Pour mener à bien notre partenariat et pouvoir se concentrer sur de nouvelles approches, il s'est rapidement avéré inévitable de faire table rase de toute la chaîne de traitement (ou *pipeline*) des données. Travail de longue haleine, à laquelle une équipe d'ingénieurs spécialistes des technologies big data s'est attelée et dont la toute première version a été livrée au bout de deux ans. Le travail a consisté essentiellement en trois points.

D'une architecture verticale vers une architecture horizontale

Dans le cadre du partenariat nous avons mis en place à la Cnam une infrastructure machine *horizontale* permettant le *calcul distribué*. Il s'agit d'une infrastructure standard aujourd'hui dans le monde du big data permettant d'éclater l'énormité des données en plusieurs paquets de beaucoup plus petite taille, pouvant être gérés par des machines aux coûts relativement modestes. Ces machines *esclaves* envoient leurs résultats à des machines *maîtres* qui les compilent et qui centralisent les résultats des calculs. Le passage à l'échelle s'effectue essentiellement par l'ajout de machines esclaves, là où l'infrastructure actuelle, centrée sur des calculateurs surpuissants (Exadata, fabriqués par Oracle) gérant seuls de très grandes quantités de données (architecture *verticale*) exigent de très gros investissements. Cette nouvelle architecture horizontale (associée à la librairie *Spark*, un autre standard du big data) permet de lire de façon efficace l'*intégralité* des données, et ceci de façon répétée, condition préalable indispensable à toute la pipeline de traitement que nous décrivons par la suite.

L'*Aplatissement* des données

Première étape indispensable de cette pipeline : la réorganisation totale des données afin de pouvoir y accéder de façon très efficace. Il s'agit, par exemple, de pouvoir récupérer rapidement tout l'historique de soin d'un assuré (les données de l'assuré sont anonymisées : ni son nom, son adresse ou son numéro de sécurité sociale ne sont accessibles). Dans l'organisation actuelle, les données d'un parcours de soin sont éclatées sur près de 800 tableaux de données (organisés dans une base propriétaire Oracle), chacune renfermant une partie du parcours. Ainsi, la table centrale (de plusieurs milliards de lignes pour une année d'historique) répertorie sur chaque ligne un remboursement particulier. Pour accéder aux détails de ce remboursement, chaque ligne va pointer vers les lignes d'autres tables (logique *relationnelle*) et ainsi de suite, reliant ainsi les 800 tables. L'*aplatissement* de ces données en

l'équivalent d'une très grosse table (distribuée sur les machines esclaves) nous permet aujourd'hui de compiler de façon très efficace tout type d'information (comme un parcours de soin).

Développement logiciel

Pour l'aplatissement des données évoqué au-dessus, et pour le développement de nouveaux modèles d'apprentissage machine, nous avons effectué de nombreux développements de codes informatiques. Ceux-ci sont essentiellement de deux types.

- Le développement d'une librairie d'interfaces normalisées permettant d'accéder à ces données non plus seulement à un niveau granulaire (données de remboursement) mais aussi au niveau des événements santé (prescriptions de médicaments, apparitions de maladies, etc.). Le but est de développer une grammaire générique facilitant (pour des non spécialistes de cette donnée et du monde du big data) la préparation des données brutes en des données numériques utilisables par les algorithmes d'apprentissage ;
- Le développement d'algorithmes d'apprentissage machine dans des cadres standards de développement utilisés couramment comme le logiciel R, Python, en se dispensant d'outils propriétaires (comme le logiciel SAS, actuellement au centre de l'analyse statistique aujourd'hui à la Cnam), très performants pour des analyses statistiques traditionnelles, mais ne permettant pas d'innovation algorithmique.

Il faut noter que la pipeline que nous venons de décrire n'a pas été encore développée pour l'entièreté de la base de données. Il était en effet nécessaire de réaliser rapidement une preuve de concept sur une partie de cette base. Ainsi, par exemple, nous n'avons pas encore exploité de données de type purement comptable, nous nous sommes concentrés exclusivement sur des informations décrivant les événements santé des assurés. Sachant que tous les outils utilisés dans ce projet sont des standards du monde du big data, le passage à l'échelle devient un problème mineur. De plus, fait notable : ils sont tous libres de droits (*open source*), donc par définition très ouverts à une recherche méthodologique sans limite.

3. Vers un algorithme de dépistage automatique de médicaments ayant des effets secondaires néfastes

Identifier un médicament sur le marché entraînant possiblement des effets secondaires néfastes est un problème très délicat. L'enjeu est bien évidemment de première importance mais l'effet secondaire, aussi grave soit-il, n'est jamais systématique et correspond à un effet de faible amplitude comparé à l'effet affiché du médicament et est donc très difficile à détecter. On parle de détection de *signaux faibles*. Pour ce faire, l'état de l'art en biostatistique consiste à faire de la *validation d'hypothèse* à l'aide de *modèles de survie*, le modèle le plus classique étant la régression de Cox [3]. Les différentes étapes de cette approche peuvent essentiellement se décomposer en 5 phases :

1. Choix d'un médicament *M* et d'un effet secondaire précis *E*. L'hypothèse que le modèle de survie cherche à valider ou à invalider est : "L'exposition d'un patient à *M* augmente-t-il le risque d'apparition de *E* ?" ;
2. Définition par un groupe de médecins experts des conditions de l'exposition : quel est le nombre minimum de prises à partir duquel le patient est considéré comme étant exposé à *M* ? Ces prises doivent-elles être rapprochées dans le temps ?
3. Extraction des données de parcours de soin de tous les assurés ayant été exposés à *M* (suivant la définition établie précédemment), on appelle une telle extraction une *cohorte* ;
4. Travail d'homogénéisation (réalisée suivant des règles établies par les mêmes experts) de la cohorte. En effet, le modèle ne permet pas, par exemple, de mettre dans un même "panier" quelqu'un qui a développé l'effet *E* un mois après l'exposition à *M* et quelqu'un qui a développé le même effet deux ans après l'exposition ;
5. Application du modèle de survie (à la cohorte ainsi constituée des patients exposées à *M*) conduisant à valider l'hypothèse avec un certain niveau de confiance. Pour cela, le modèle s'appuie sur une comparaison entre la sous-cohorte des patients ayant développé *E* et la sous-cohorte *contrôle* des patients n'ayant pas développé *E*.

Ce processus très éprouvé continue largement à faire ses preuves. Il a cependant deux inconvénients majeurs : il faut définir a priori *M* et *E*; les étapes 2. et 4. sont le fruit de discussions délicates d'experts rendant le processus complet relativement long (il faut compter classiquement de l'ordre de plusieurs mois pour avoir des réponses fiables). Cela rend totalement prohibitif une analyse systématique à grande échelle sur un grand nombre de médicaments et/ou d'effets secondaires.

C'est dans ce cadre que s'inscrit le travail des trois premières années du partenariat : l'élaboration d'un algorithme de dépistage automatique des médicaments augmentant les risques d'un effet secondaire donné, algorithme ne nécessitant qu'une intervention ponctuelle d'experts, réduisant les étapes 2. et 4. précédemment décrites à leur strict minimum. Bien entendu, on ne peut espérer qu'un tel algorithme soit aussi précis qu'une analyse classique avec modèle de survie spécifique à un couple (*M*, *E*) précis. D'une vitesse d'élaboration sans commune mesure, il faudra le considérer comme une étape préalable permettant d'identifier, sur un grand nombre de médicaments, ceux susceptibles d'être problématiques. Les médicaments ainsi identifiés nécessiteront une analyse plus approfondie pour confirmation. Cet algorithme de dépistage intervient donc en amont des analyses classiques (de type survie) et permettra un premier dépistage rapide à grande échelle.

ConvSCCS : Un algorithme de dépistage à grande échelle

L'algorithme se base sur le principe de Self-Controlled Case Series (SCCS) [8, 10]. On ne retient pour l'étude que les patients ayant développé l'effet secondaire étudié (d'où le nom "case-series" dans SCCS). Le *contrôle* ne se fait plus alors avec une cohorte de patients n'ayant pas contracté l'effet secondaire, mais avec les patients l'ayant contracté avant que l'effet n'apparaisse. Chaque patient joue ainsi le rôle de son propre contrôle (c'est la signification du "self-controlled" dans SCCS) : la période où il contracte l'effet secondaire

est comparée implicitement à celle où il ne l'a pas encore contracté bien qu'il soit déjà exposé aux médicaments à dépister. L'utilisation de ce type de modèle est une alternative intéressante au modèle des risques proportionnels de Cox pour l'identification de médicaments ayant des effets secondaires néfastes. Elle présente essentiellement deux avantages par rapport à ce type de modèle :

- Elle améliore significativement le "niveau" des signaux à détecter, les signaux sont alors "moins faibles" et donc plus facile à détecter qu'avec un modèle de survie de type régression de Cox. Cela vient du fait que le nombre d'assurés ayant développé l'effet secondaire est beaucoup plus petit que le nombre d'assurés exposés aux médicaments considérés, problème qui s'estompe lorsque l'on ne considère plus de sous-cohorte contrôle ;
- Elle permet de diminuer sensiblement l'effet de biais potentiellement présents dans les données. Ce type de modèle n'est en effet sensible qu'aux variations des variables *longitudinales* considérées (des valeurs qui changent le long de la période de temps considérée), comme les expositions médicamenteuses. Des variables statiques, telles que le genre du patient, n'impactent pas les résultats d'un tel modèle ;

Cependant, les algorithmes standards de type SCCS sont loin d'être exempts de défauts et notre algorithme ConvSCCS, introduit dans [5], présente des améliorations très significatives, particulièrement adapté pour du dépistage automatique à grande échelle :

- ConvSCCS permet de modéliser l'effet potentiel de plusieurs médicaments simultanément, alors que les méthodes SCCS classiques ne permettent d'étudier qu'un seul médicament à la fois. ConvSCCS est donc bien moins sensible aux effets de *confusion* entre médicaments. En effet, un patient est exposé dans son parcours de soin à plusieurs médicaments, et ces expositions sont souvent superposées. Lorsque l'on n'incorpore pas assez de médicaments dans la modélisation, le modèle a alors tendance à confondre l'effet d'un médicament avec l'effet de ceux non utilisés dans la modélisation ;
- L'algorithme ConvSCCS a de bonnes propriétés de robustesse permettant de diminuer la sensibilité des résultats obtenus par le modèle au travail préliminaire des experts (points 2. et 4. ci-dessus). Les étapes 2. et 4. peuvent alors se réduire à un travail relativement sommaire ;
- ConvSCCS permet d'obtenir des courbes quantifiant l'influence de l'exposition à chaque médicament en fonction du temps sur la probabilité d'apparition de l'effet secondaire. Cet algorithme apprend donc automatiquement les durées d'exposition amenant éventuellement à un risque d'apparition de l'effet secondaire, ce qui n'était pas le cas des autres méthodes SCCS. Pour cela, nous utilisons une technique dite de *pénalisation*, qui va forcer les coefficients appris par le modèle à avoir une structure "simple", tout en expliquant correctement les données. On observe en effet dans la Figure 1 plus bas des courbes constantes par morceaux, sur des intervalles relativement longs. Cette forme de courbe est précisément celle recherchée ici : nous voulons détecter des changements de valeurs statistiquement significatifs, qui s'interprètent comme des changements importants de l'influence de l'exposition, après une certaine durée, sur la probabilité d'occurrence de l'effet secondaire. Par ailleurs,

ces courbes sont calculées avec une estimation de l'incertitude d'estimation de ces influences.

Premiers résultats obtenus avec l'algorithme ConvSCCS : Identification d'un anti-diabétique augmentant le risque du cancer de la vessie

Les premiers résultats obtenus avec ConvSCCS sont liés au projet "pilote" du partenariat. L'objectif était la détection "en aveugle" du pioglitazone, dont un effet secondaire de sur-risque de cancer de la vessie a été confirmé dans [6] entraînant son retrait du marché en 2011. Cet exemple précis a été choisi de concert avec les médecins en santé publique de la Cnam, car il s'agit d'un signal considéré comme faible. Le pioglitazone a en effet été retiré du marché en France [6], mais pas partout dans le monde [4]. Ce projet pilote a donc consisté à retrouver cet effet déjà connu par la Cnam, afin de valider deux aspects de notre approche :

- la pipeline (qui remet à plat l'ensemble des traitements de données) en reproduisant précisément les résultats obtenus dans [6] avec des régression de Cox;
- la validité de l'algorithme ConvSCCS, comme algorithme alternatif permettant de retrouver l'effet de cette molécule en aveugle, en simplifiant grandement les étapes 1 à 5 décrites au-dessus, et qui apporte un grand nombre d'améliorations citées au-dessus.

L'étude est basée sur une cohorte de 2,5 millions de patients diabétiques de type 2 sur 4 ans d'historique, réparties sur environ 2 milliards de lignes (1,3 téraoctets). L'effet indésirable considéré est donc le cancer de la vessie, et nous considérons des expositions à des antidiabétiques listés dans la Figure 1. Sur ce jeu de données, les étapes d'aplatissement et de préparation des données demandent environ 40 minutes de calcul. L'entraînement du modèle (phase d'apprentissage machine) dure quelques minutes seulement (effectué avec notre librairie tick [1]). Nous illustrons dans la Figure 1 les courbes obtenues par le modèle. Ces courbes quantifient l'impact des expositions, dans le temps, à des molécules antidiabétiques sur le risque de cancer de la vessie. La valeur 1 correspond à une absence d'effet (dans les deux sens) : autour de cette valeur, la molécule n'a pas d'impact sur le risque. On observe que seul le pioglitazone a un effet significativement supérieur à 1, et ce après une exposition d'un peu plus d'un an. Ce résultat est en parfaite cohérence avec les résultats de [6] qui avaient été obtenus en suivant un protocole biostatistique très précis (voir points 1 à 5 au-dessus) de validation d'hypothèse, tandis que le résultat obtenu dans la Figure 1 a été obtenu avec une approche beaucoup plus automatisable et facile à déployer à grande échelle (grand nombre de médicaments et grand nombre d'effets indésirables).

4. Travail en cours et perspectives

Les projets d'étude du nouveau partenariat

L'algorithme ConvSCCS [5], même dans sa toute première version, est assez mûr pour être maintenant testé en "situation réelle". C'est ce à quoi nous nous employons aujourd'hui.

Nous travaillons sur les médicaments actuellement sur le marché qui augmenteraient le risque de chute chez les personnes âgées. Les experts de la Cnam ont identifié près de 250 médicaments (150 molécules) à étudier parmi les antihypertenseurs, les antidépresseurs, les neuroleptiques et les hypnotiques. La cohorte qui nous intéresse contient près de 12 millions de personnes répartie sur environ 2 milliards de lignes par an. L'étape d'aplatissement et de préparation de la donnée prend moins de 2h et l'entraînement du modèle dure toujours quelques minutes seulement.
Il s'agit d'un véritable passage à l'échelle (facteur de l'ordre de 4) par rapport au projet pilote sur les antidiabétiques. Passage à l'échelle pour l'ensemble de la pipeline : infrastructure machine et logicielle, mais aussi de l'algorithme lui-même qui supporte une plus grande quantité de données et une robustesse par rapport à ce changement de problématique. En effet, le projet pilote concerne des effets à long-terme (au moins un an d'attente avant de voir l'effet de l'exposition au pioglitazone) alors que dans le cas des chutes, l'effet, s'il a lieu, est à court-terme (quelques jours). Nous travaillons également sur d'autres améliorations importantes, notamment la possibilité de prendre en compte simultanément plusieurs effets secondaires, voire de détecter automatiquement ce que sont les effets secondaires de tel ou tel médicament.
En parallèle de ce travail de recherche en pharmacovigilance et de sa "mise en production", deux autres thèmes seront abordés lors du partenariat : la lutte contre la fraude et l'identification de facteurs utiles à l'analyse des parcours de soins. Nous avons entamé un travail de visualisation interactive d'un grand nombre de parcours de soins, étape essentielle pour appréhender les parcours au sein d'une pathologie donnée. Nous travaillons également sur d'autres techniques d'apprentissage machine, notamment l'apprentissage profond (deep learning) ainsi que d'autres techniques d'intelligence artificielle.

Vers une systématisation et une ouverture de la pipeline big data

Tous les développements qui ont été réalisés dans le cadre de ce partenariat ont été faits avec un objectif d'utilisation par des non-experts en big-data. Le transfert de connaissance a déjà commencé via une mutualisation des développeurs informatiques. Notre ambition est d'être en particulier une preuve de concept pour une ouverture de la pipeline à plus grande échelle (tous les logiciels que nous avons développés sont libres de droits), une plateforme alternative à celle actuellement présente dans le SNDS qui définit aujourd'hui le cadre légal d'accès au Sniiram, une preuve de concept pour le *hub des données de santé* annoncé par le Président de la République lors de la remise du rapport Villani le 29 Mars 2018. Un accès moderne à ce jeu de données unique au monde source de tant de richesses encore très largement sous-exploitée pour le bien commun.

5. Remerciements

Ce partenariat est avant tout un travail d'équipe. Les résultats sont le fruit de la collaboration d'équipes fortement multidisciplinaires. Nous tenons à remercier les médecins, développeurs, chercheurs et experts-métiers de la Cnam, en particulier Aurélie Bannay, Hélène Caillol, Joël

Coste, Claude Gissot, Anke Neumann, Jérémie Rudant et Alain Weill et les équipes de développeurs, data-scientists et chercheurs de l'Ecole Polytechnique, en particulier Prosper Burq, Philip Deegan, Xristos Giastidis, Agathe Guilloux, Daniel de Paula da Silva et Youcef Sebiat.

RÉFÉRENCES

[1] Bacry E, Bompaire M., Gaïffas *et al.* Tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. Journal of Machine Learning Research. 2018.
[2] Coloma P. Mining Electronic Healthcare Record Databases to Augment Drug Safety Surveillance. PhD Manuscript, University Medical Center Rotterdam. 2012.
[3] Cox D. Regression Models and Life-Tables. Journal of the Royal Statistical Society, Series B. 1972.
[4] Lewis J, Habel L, Quesenberry C *et al.* Pioglitazone use and risk of bladder cancer and other common cancers in persons with diabetes. JAMA. 2015.
[5] Morel M, Bacry E, Gaïffas S *et al.* ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection. ArXiv preprint. 2017.
[6] Neumann A, Weill A, Ricordeau P *et al.* Pioglitazone and risk of bladder cancer among diabetic patients in France : a population-based cohort study. Diabetologia. 2012.
[7] Rajkomar A, Oren E, Chen, K, *et al.* Scalable and accurate deep learning for electronic health records. arXiv preprint. 2018.
[8] Schuemie M, Trifirò G, Coloma P *et al.* Detecting adverse drug reactions following long-term exposure in longitudinal observational data: The exposure-adjusted self-controlled case series. Statistical methods in medical research. 2014.
[9] Shickel B, Tighe P, Bihorac A, *et al.* Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. ArXiv e-prints. 2018.
[10] Whitaker H, Paddy Farrington C, Spiessens B *et al.* Tutorial in biostatistics: the self- controlled case series method. Statistics in medicine. 2006.

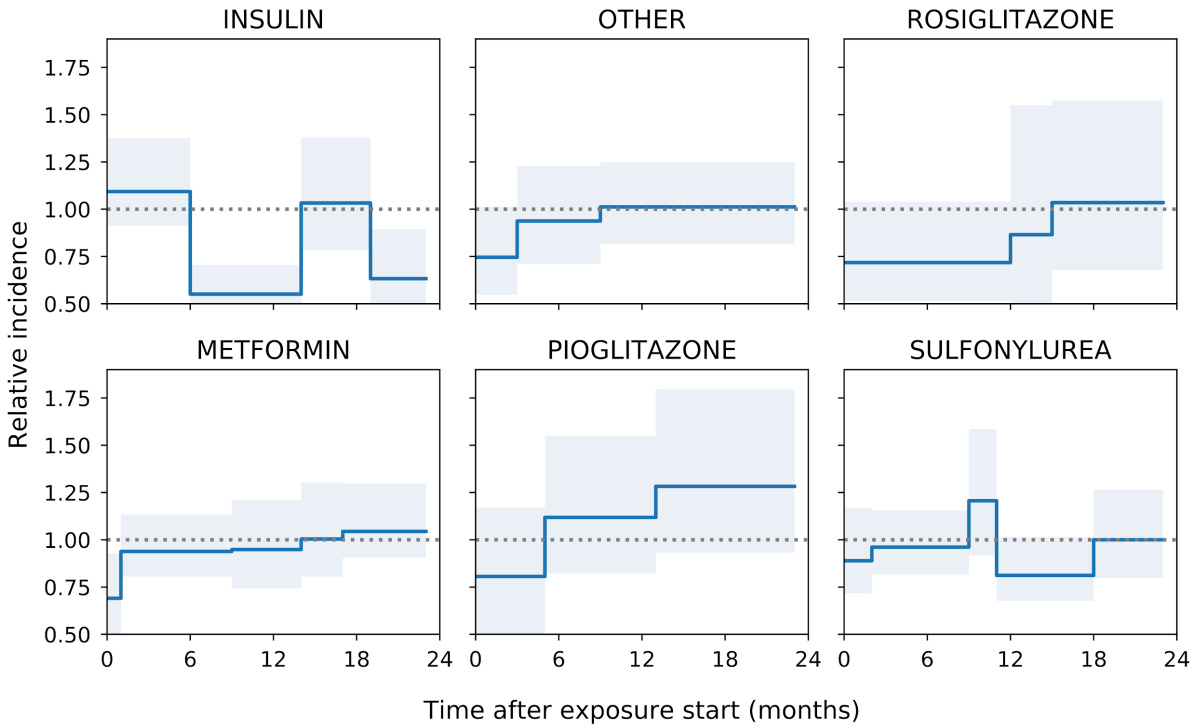

Figure 1. Estimation des effets des expositions

Estimations des effets de l'exposition à différents antidiabétiques sur le risque d'apparition du cancer de la vessie. On observe que seul le pioglitazone a un effet significativement supérieur à 1.