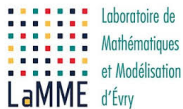


# Machine en learning en santé : exemple en pharmacovigilance

Agathe Guilloux  
LaMME - Université d'Évry - Paris Saclay



Contexte

# Données médico-administratives : SNDS<sup>1</sup>/ SNIIRAM<sup>2</sup> & PMSI<sup>3</sup>

- ▶ Données de parcours de soin liées aux remboursements
- ▶ 99% de la population française est couverte depuis 2002

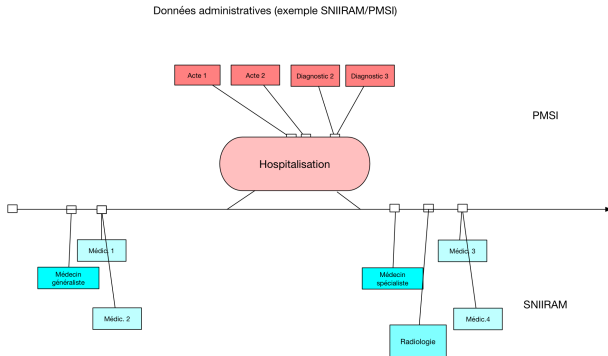


Figure 1: Schéma pour un patient

<sup>1</sup>Système national des données de santé

<sup>2</sup>Système national d'information inter-régimes de l'Assurance maladie

<sup>3</sup>Programme de médicalisation des systèmes d'information

# Données de soins hospitaliers

## ► Données d'entrepôt (ex : HEGP<sup>4</sup>)

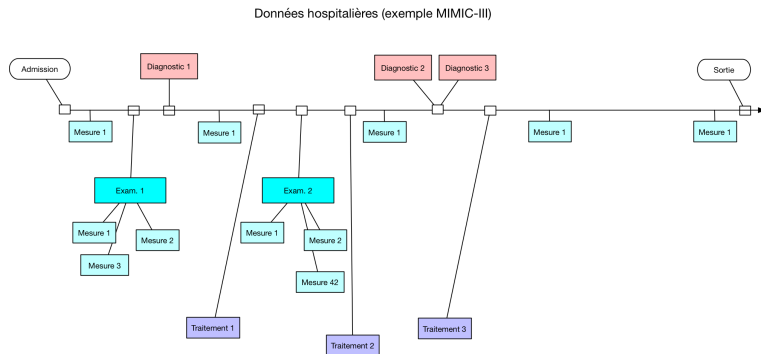


Figure 2: Schéma pour un patient

# Particularités

## Données médico-administratives

- ▶ ne contiennent pas de données cliniques, génomiques, etc.. (pour l'instant<sup>5</sup>)
- ▶ l'obligation de non-réidentification ne permet pas de suivi long

## Données de soins hospitaliers

- ▶ la structure dépend de l'hôpital/groupement d'hôpitaux (pour l'instant<sup>6</sup>)

## En général

- ▶ temporalité
- ▶ complexité
- ▶ taille

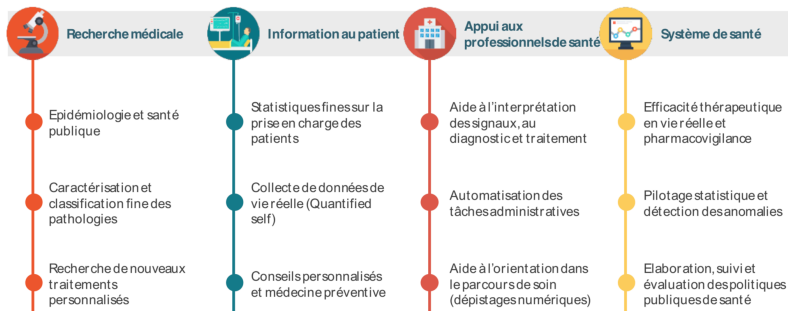
---

<sup>5</sup>"La France [...] doit créer une plate-forme d'accès et de mutualisation des données pertinentes pour la recherche et l'innovation en santé", Villani 2018

<sup>6</sup>" Il faudra assurer l'interopérabilité des données et des systèmes (des hôpitaux jusqu'au SNDS)", Villani 2018

# Futur health data hub<sup>7</sup>

- ▶ “besoin d'un guichet unique, assurant un accès simplifié, effectif et accéléré aux données”
- ▶ “les données de santé financées par la solidarité nationale constituent un patrimoine commun”
- ▶ “Ces données doivent donc être mises [...] au service du plus grand nombre dans le respect de l'éthique et des droits fondamentaux de nos concitoyen”



<sup>7</sup>Cuggia, Polton, and Wainrib 2018

## Exemples de questions pour le machine learning

- ▶ **En pharmacovigilance** : relier la prise de médicaments avec des effets indésirables (adverse events)
- ▶ **Administratives** : prédire la durée de séjour à partir des 1eres heures, codage automatique d'actes
- ▶ **En clinique** : prédire les re-hospitalisations précoces

Partenariat Ecole polytechnique/CNAMTS (2015-2020)

## Cancer de la vessie pour une cohorte de diabétiques

Le “projet pilote”<sup>8</sup> en pharmaco-vigilance sur les données du SNIIRAM<sup>9</sup>/PMSI<sup>10</sup>

### But

développer une méthode de “screening” permettant un premier balayage automatique sur plusieurs médicaments.

- ▶ ≠ validation d'hypothèse
- ▶ Etape simplifiée de préparation de cohorte

### Application

- ▶ Cohorte : diabétiques de type 2
- ▶ Classe pharmaceutique : anti-diabétique
- ▶ Effet indésirable : cancer de la vessie

→ identification du Pioglitazone (retiré du marché en 2011, Neumann et al. 2012)

---

<sup>8</sup>L'équipe : E. Bacry , S. Gaïffas, A. Guilloux, M. Morel, F. Leroy, ...

<sup>9</sup>Système national d'information inter-régimes de l'Assurance maladie

<sup>10</sup>Programme de médicalisation des systèmes d'information

# Cancer de la vessie pour une cohorte de diabétiques

## Quelques chiffres

- ▶ 2.5 millions de personnes
- ▶ 4 ans d'historique
- ▶ 1.3 To ( $\simeq$  250Go par an)
- ▶ 2 milliards de "lignes" (500 millions de lignes par an)
- ▶ "Applatissage de la base"  $\simeq$  40 minutes (depuis spark)

## Résultats obtenus par un algorithme “standard”

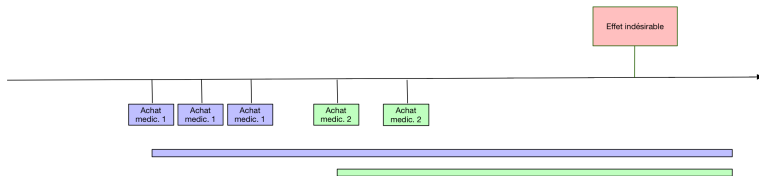
### **Modèle de Cox (survie) ⇒ validation d'hypothèse**

- ▶ Papier de référence : *Pioglitazone and risk of bladder cancer among diabetic patients in France : a population-based cohort study*, A.Neumann, A.Weill, P.Ricordeau, J.P.Fagot, F.Alla - H.Allemand Diabetologia 2012.

### **But de l'étude :**

- ▶ Reproduction des résultats
- ▶ Etude de la stabilité par rapport aux paramètres de construction de cohorte

# Exposition et effet indésirable en pharmacovigilance



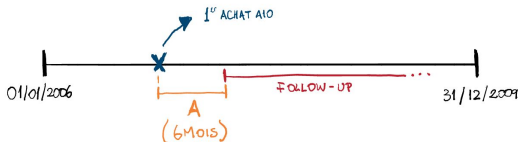
# Les paramètres de construction de la cohorte

## 1. Définition du cancer

- ▶ “Broad” : diagnostic C67
- ▶ “Narrow” : C67 + radiothérapie + chimiothérapie + ...

## 2. Période de controle du “Follow up”

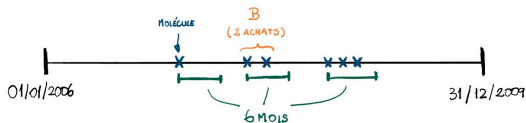
- ▶ Patients dont le cancer est diagnostiqué avant ou pendant la période A sont écartés
- ▶ Défaut : 6 mois



# Les paramètres de construction de la cohorte (suite)

## 3. Minimum d'achats

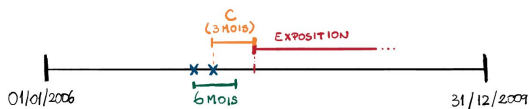
- ▶ Un patient est considéré comme exposé à une molécule s'il a acheté un médicament contenant cette molécule  $B$  fois en  $N$  mois.
- ▶ Défaut :  $B = 2$ ,  $N = 6$  mois



# Les paramètres de construction de la cohorte (suite et fin)

## 4. Délai d'exposition

- ▶ Nombre de mois après la dernière prise de médicament pour que le patient soit considéré comme exposé.
- ▶ Défaut : 3 mois



Un nouveau modèle pour la pharmacovigilance

## Un nouveau modèle de screening

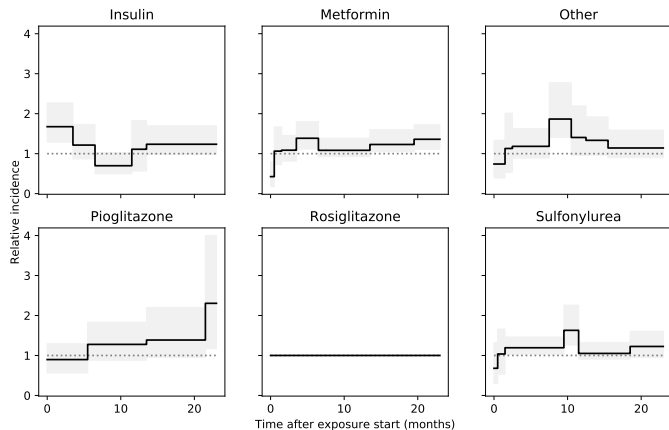
Création d'un nouveau modèle (in Morel et al. 2019) :

- ▶ Un modèle "self-controlled case-series" (SCCS) : on ne garde dans la cohorte que les personnes ayant eu un cancer
- ▶ Patients forment leur propre contrôle
- ▶ Travail de préparation de cohorte très simplifiée : définition de l'exposition
- ▶ On estime l'impact (en fonction du temps) de l'exposition à un médicament sur la probabilité de développer un cancer à tel moment

Un peu plus spécifiquement

- ▶ Structure auto-régressive des features
- ▶ Techniques de pénalisation et cross-validation (signal faible)
- ▶ Algorithmes d'optimisation extrêmement rapides

## Résultats obtenus par le nouveau modèle



Incidences relatives sur le risque de cancer de la vessie

# Limites du modèle et perspectives

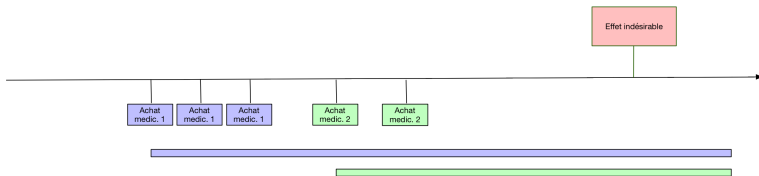
## Avantages

- ▶ **Rapidité** : quelques secondes pour entraîner le modèle (après flattening et featurizing)
- ▶ **Scalabilité** sur une machine à de grosses cohortes
- ▶ **Plusieurs expositions** : solution à un problème majeur des modèles SCCS, limité à l'exposition à **une seule** molécule dans la littérature !

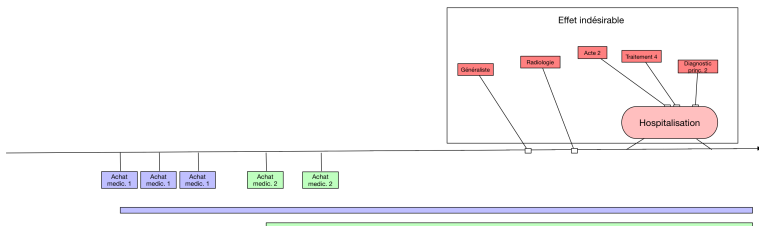
## Les problèmes qui restent

- ▶ L'hypothèse d'exogénéité des features longitudinales est très forte en pharmaco-vigilance
- ▶ La définition des effets indésirables est encore lourde, il faudrait l'automatiser...

## Exposition et effet indésirable en pharmacovigilance (2)



En fait, plutôt



## Exemples de questions pour le machine learning

- ▶ **En pharmacovigilance** : relier la prise de médicaments avec des effets indésirables (adverse events)
- ▶ **Administratives** : prédire la durée de séjour à partir des 1eres heures, codage automatique d'actes
- ▶ **En clinique** : prédire les re-hospitalisations précoces

## References I



M. Cuggia, D. Polton, and G. Wainrib. *Mission de préfiguration pour le health data hub*. 2018. URL: [https://solidarites-sante.gouv.fr/IMG/pdf/181012\\_-\\_rapport\\_health\\_data\\_hub.pdf](https://solidarites-sante.gouv.fr/IMG/pdf/181012_-_rapport_health_data_hub.pdf).



Maryan Morel et al. *ConvSCCS: convolutional self-controlled case series model for lagged adverse event detection*. 2019. eprint: [arXiv:1712.08243](https://arxiv.org/abs/1712.08243).



A Neumann et al. “Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study”. In: *Diabetologia* 55.7 (2012), pp. 1953–1962.



C. Villani. *Donner un sens à l'intelligence artificielle*. 2018. URL: <https://www.aiforhumanity.fr>.