

Les grandes banques de données médicales et administratives du Québec pour la recherche en santé



Université de
Sherbrooke

Yohann M Chiu, PhD

Département de médecine de
famille et médecine d'urgence

29 mai 2025

Présentation à l'EINS 2025



Les grandes banques de données médicales et administratives du Québec pour la recherche en santé



Yohann M Chiu, PhD
Département de médecine de
famille et médecine d'urgence

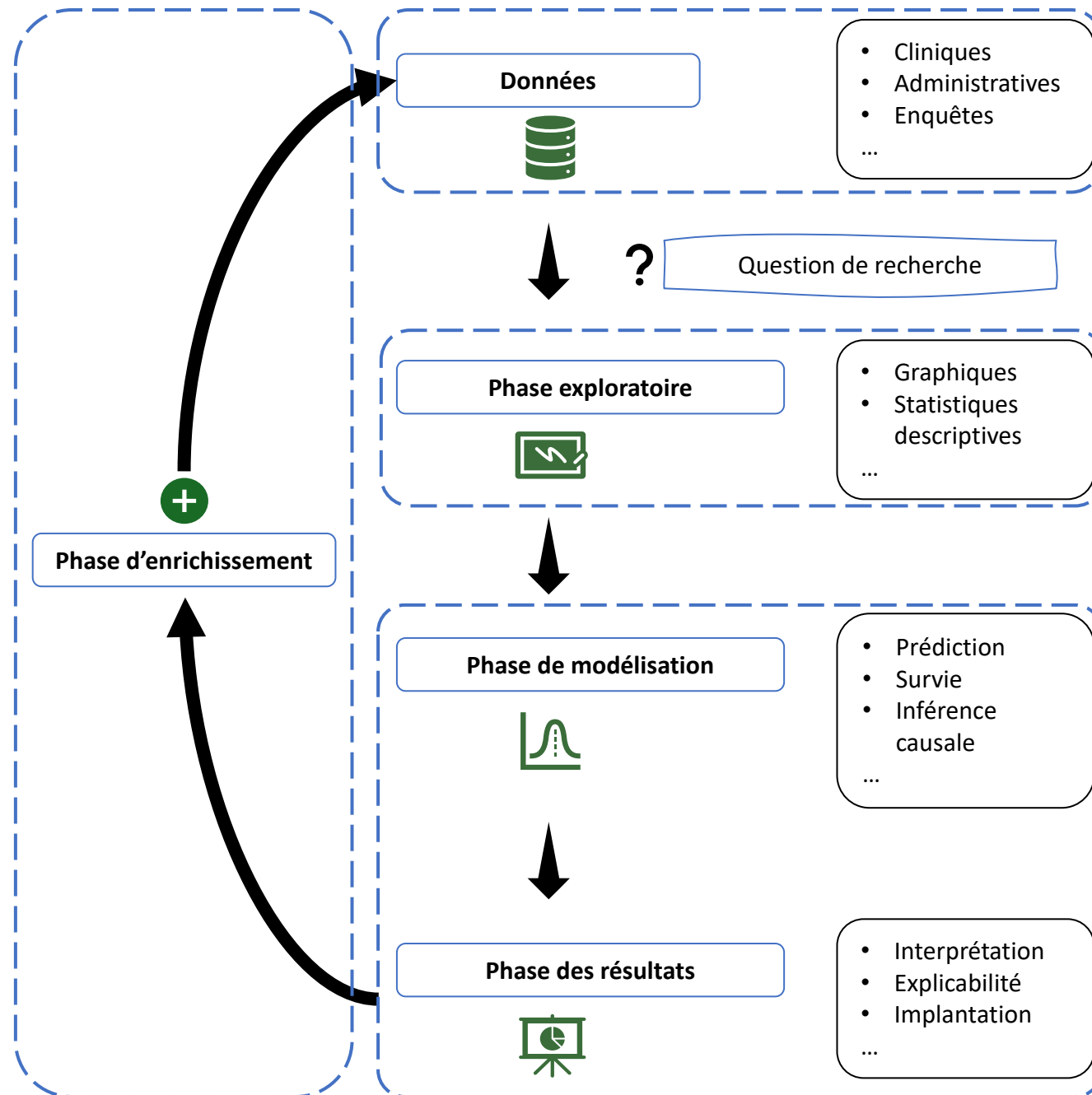
29 mai 2025
Présentation à l'EINS 2025



Conflits d'intérêt

- Pour les travaux présentés: CRCHUS, FRQS, IRSC
- Aucun autre conflit à signaler





I. Les banques de données médico-administratives pour la recherche

- Avantages et inconvénients
- SISMACQ

II. Grande utilisation des services d'urgence

- Profils des grands utilisateurs
- Construction d'un outil de dépistage

III. Surveillance de la polypharmacie

- Fouille de données
- Prédiction
- Transparence

IV. Conclusions et réflexions

- On n'a pas parlé de...
- Statistique et IA pour la recherche en santé

À propos	
Banques de données ministérielles	
APR-DRG	
BDCU	
CDLAB	
ICLSC	
MED-ECHO	
PIJ	
RQC	
SIRTQ	
Autres sources de données	

Cette section regroupe les différentes banques de données ministérielles. Pour chacune d'entre elles, vous retrouverez une fiche descriptive qui indique notamment le nom de l'organisme propriétaire, la description sommaire de la banque de données ainsi que sa ou ses sources d'alimentation et ses finalités.

Données administratives Données d'enquêtes

Pour en savoir plus sur les données administratives disponibles, vous pouvez effectuer une recherche par mot-clé et/ou filtrer les données par secteur d'activité et organisme détenteur.

Mot clé

Secteur d'activité

Organisme détenteur

Ministère de la Santé et des Services sociaux (MSSS)
Registre des événements démographiques (RED)

Ministère de l'Éducation - Ministère de l'Enseignement supérieur (MEQ-MES)

Revenu Québec (RQ)
Particuliers

Régie de l'assurance maladie du Québec (RAMQ)

[Fichier d'inscription des personnes assurées \(FIPA\)](#)

[Services pharmaceutiques \(MED\)](#)

[Services médicaux rémunérés à l'acte \(MOD\)](#)

[Banque de données communes des urgences \(BDCU\)](#)

[Maintenance et exploitation des données pour l'étude de la clientèle hospitalière \(MED-ECHO\)](#)

<https://statistique.quebec.ca/services-recherche/donnees/administratives>

I – Avantages et inconvénients

Utilisation secondaire de données clinico-pharmaco-

+ Design alternatif pour l'utilisation, l'efficacité

+ Évaluer les traitements dans des **contextes réels** observationnelles

- Moins de contrôle sur la qualité d'évaluation de la recherche

- Banques de données **administratives**

ACCESSIBILITÉ

Data Sources for Pharmacoepidemiology and

H.

[/doi.org/10.1592/phco.29.2.138](https://doi.org/10.1592/phco.29.2.138) | Citations: 54



PDF



TOOLS



SHARE

Journal of Cardiology

, March–April 2012, Pages 162-168



and Surveillance of Disease: Can We Trust Administrative Hospital Data?

[Claudia Blais PhD^{b,c}](#), [Denis Hamel MSc^b](#), [Kevin Brown MSc^a](#),

[Raymond Cartier MD^a](#), [Maude Giguère^a](#), [Céline Carroll BSc^a](#),

[Peter Bogaty MD^d](#)

I – Les BD du Québec

Banques de données créées à l'origine pour le remboursement des actes médicaux

1. MED-ÉCHO: séjour hospitalier, diagnostics, services...
2. FIPA: âge, régime public d'assurance médicaments...
3. Services pharmaceutiques: médicament, dose, durée...
4. Services médicaux: acte médical, diagnostics...
5. RED: date, cause principale...

Services
pharmaceutiques
Médicaments

Services
médicaux
Médecins

MED-ÉCHO
Hospitalisations

FIPA
Assurance
médicaments

RED
Décès

I – Les BD du Québec

Exemple des services pharmaceutiques

Banque de données **massives** (à valider avec Pre Khnaisser)

Une seule table contenant des renseignements sur les services pharmaceutiques

Alerte au biais: **personnes assurées** au public uniquement

	noindiv_srap	SMED_pgm_med	SMED_cod_plan	dateService	SMED_cod_DIN	classe_AHFS
1	A1000001	01	96	2013-10-22	10000003	100005
2	A1000002	AD	2K	2015-11-14	10000001	100001
3	A1000003	01	20	2012-10-09	10000005	100005
4	A1000004	02	11	2011-01-06	10000002	100009
5	A1000005	01	2S	2014-04-21	10000000	100008
6	A1000006	AL	1L	2015-01-13	10000006	100008
	SMED_cod_denom_comne	SMED_cod_forme_med	SMED_cod_tenr_MED	SMED_cod_nat_expr_ordnc		
1	10003	10001	10004			RS
2	10007	10008	10005			NS
3	10005	10009	10004			NV
4	10008	10000	10008			NS
5	10005	10009	10000			NS
6	10008	10001	10009			NS
	SMED_cod_selec_med	SMED_NB_jr_duree_trait	SMED_qte_med	cout_brut	contr_pers_assur	
1	P	29	85.22396	144.73514	99.68627	
2	E	14	13.93031	119.01956	93.29204	
3	C	15	55.02035	43.05313	93.82131	
4	E	29	91.58785	197.15901	57.42764	
5	B	3	27.40131	64.24019	99.09444	
6	E	19	22.60442	78.91139	76.34049	
	SMED_classe_presc	no_banal_presc	specPres			
1	7	1000000078	12			
2	7	1000000031	15			
3	7	1000000010	11			
4	4	1000000064	14			
5	5	1000000049	14			
6	2	1000000035	14			

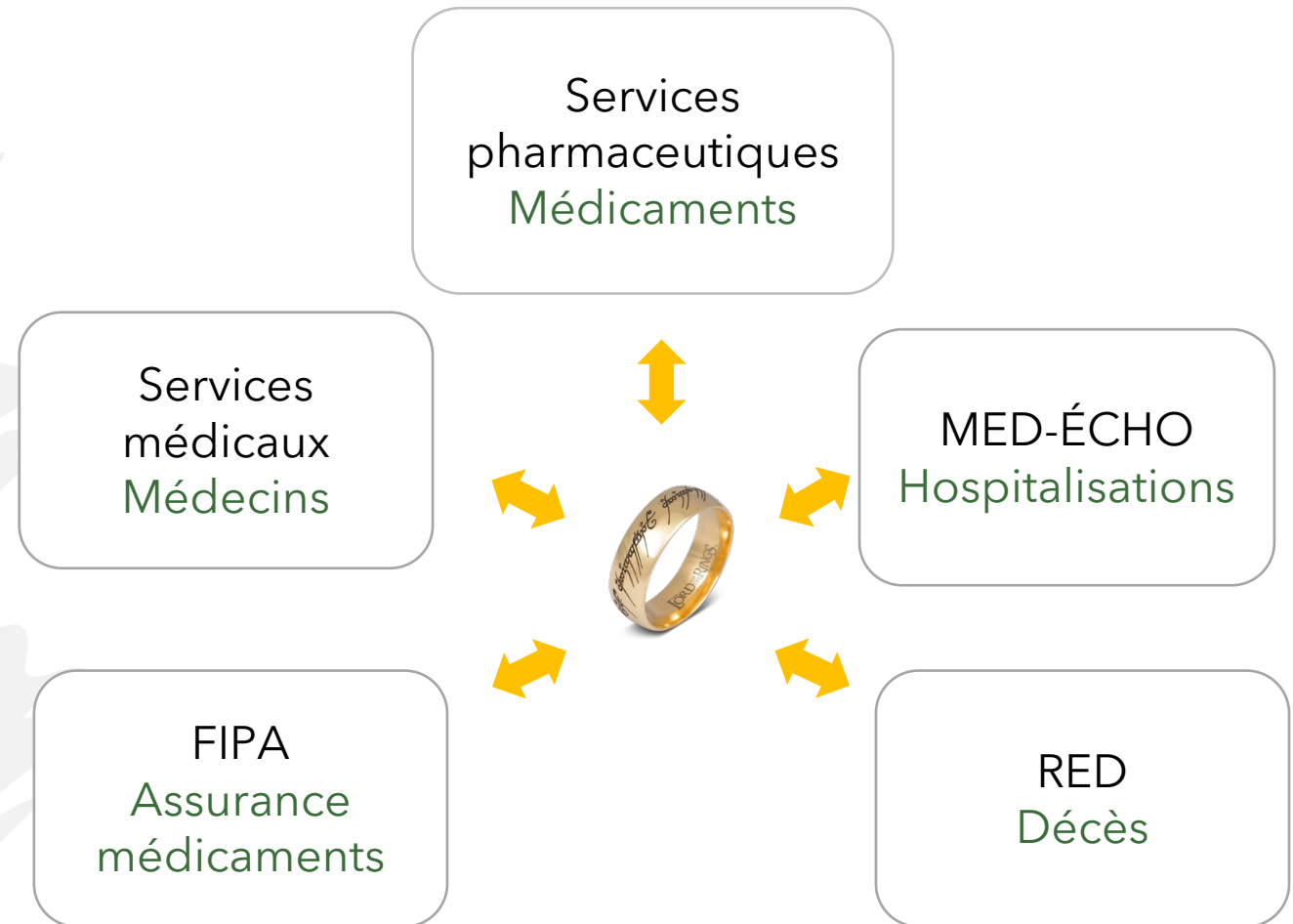
I – SISMACQ

Jumelage par un identifiant unique

➔ Système intégré de surveillance des maladies chroniques du Québec

Hébergé dans un environnement sécurisé à l'INSPQ

Études observationnelles ou **surveillance populationnelle** de maladies chroniques



I – SISMACQ

Jumelage par un identifiant unique

➔ Système intégré de surveillance des maladies chroniques du Québec

Hébergé dans un environnement sécurisé à l'INSPQ

Études observationnelles ou **surveillance populationnelle** de maladies chroniques



Maladie	Âge (ans)	Définition de cas	Code diagnostique	
			CIM-9	CIM-10-CA
Cardiopathies ischémiques	20 et plus	Deux diagnostics dans les services médicaux en 1 an OU un diagnostic (principal ou secondaire) dans MED-ÉCHO	410 à 414	I20 à I25
Diabète	1 et plus (diabète)	Deux diagnostics dans les services médicaux en 2 ans OU un diagnostic (principal ou secondaire) dans MED-ÉCHO	250	E10 à E14 (diabète)

- I. Les banques de données médico-administratives pour la recherche
 - Avantages et inconvénients
 - SISMACQ

- II. Grande utilisation des services d'urgence
 - Profils des grands utilisateurs
 - Construction d'un outil de dépistage

- III. Surveillance de la polypharmacie
 - Fouille de données
 - Prédiction
 - Transparence

- IV. Conclusions et réflexions
 - On n'a pas parlé de...
 - Statistique et IA pour la recherche en santé

II – Grande utilisation des SU

≥4 visites à l'urgence par année¹

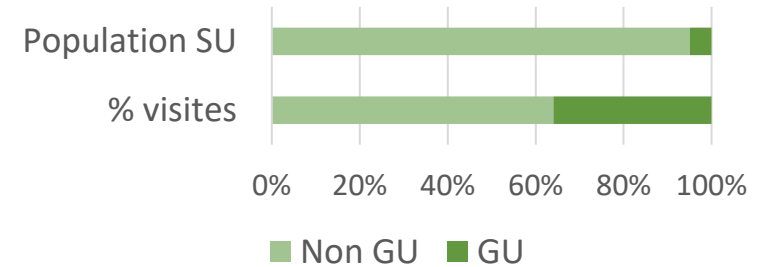
- ➔ soins fragmentés, épisodiques
- ➔ qualité sous-optimale

Gestion de cas jugée efficace pour prévenir la grande utilisation²

Prédiction de la grande utilisation ~3%

➔ « débalancement »³

1. Krieg, C., et al. (2016). Individual predictors of frequent emergency department use: a scoping review. *BMC health services research*, 16, 1-10.
2. Hudon, C., et al. (2016). Effectiveness of case management interventions for frequent users of healthcare services: a scoping review. *BMJ open*, 6(9), e012353.
3. Chiu, Y. M., et al. (2023). Machine learning to improve frequent emergency department use prediction: a retrospective cohort study. *Scientific Reports*, 13(1), 1981.



[nature](#) > [scientific reports](#) > [articles](#) > article

Article | [Open access](#) | Published: 03 February 2023

Machine learning to improve frequent emergency department use prediction: a retrospective cohort study

[Yohann M. Chiu](#) , [Josiane Courteau](#), [Isabelle Dufour](#), [Alain Vanasse](#) & [Catherine Hudon](#)

II – Grande utilisation des SU

Nécessité de bien comprendre les besoins car profils hétérogènes

Méthodes d'apprentissage non supervisé pourrait nous aider à définir des profils cliniquement significatifs?

Méthodes d'apprentissage supervisé pourrait nous aider à définir les facteurs prédictifs de la grande utilisation?



Variable		Total (%)	Non grands utilisateurs (%)	Grands utilisateurs (%)
Total		451,775 (100)	438,099 (100)	13,676 (100)
Femme		234,320 (51.9)	226,968 (51.8)	7,352 (53.8)
Âge	18-34	23,723 (5.3)	22,775 (5.2)	948 (6.9)
	35-54	83,393 (18.5)	80,977 (18.5)	2,416 (17.7)
	55-64	99,136 (21.9)	96,618 (22.1)	2,518 (18.4)
	65-74	116,323 (25.7)	113,198 (25.8)	3,125 (22.9)
	75-84	93,091 (20.6)	89,887 (20.5)	3,204 (23.4)
	≥ 85	36,109 (8.0)	34,644 (7.9)	1,465 (10.7)
Charlson	0	277,798 (61.5)	272,919 (62.3)	4,879 (35.7)
	1-2	98,228 (21.7)	94,558 (21.6)	3,670 (26.8)
	3-4	34,395 (7.6)	32,248 (7.4)	2,147 (15.7)
	≥ 5	41,354 (9.2)	38,374 (8.8)	2,980 (21.8)
Douleurs chroniques		75,263 (16.7)	71,859 (16.4)	3,404 (24.9)
Troubles mentaux		15,778 (3.5)	14,281 (3.3)	1,497 (10.9)
Hospitalisation		191,862 (42.5)	182,364 (41.6)	9,498 (69.5)
Hautement défavorisés (social)		103,955 (23.0)	100,059 (22.8)	3896 (28.5)

II – Grande utilisation des SU

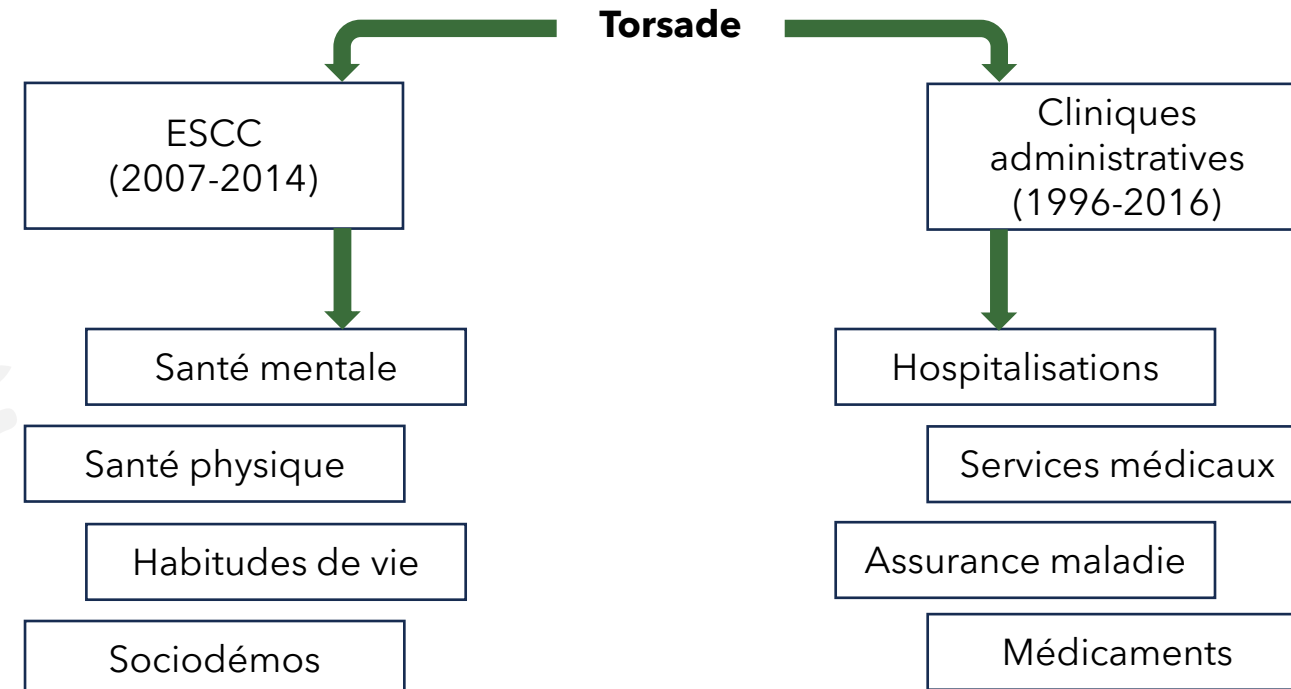
À vous!

- Explorer les données (dictionnaire + BD)
- Décrire les données selon sexe, âge, indice de comorbidité, présence de douleur, santé perçue, santé mentale perçue, région...
- Créer un indicateur de décès
- Description selon le statut de grande utilisation (>4 visites dans l'année précédente)

Cohort Profile

Cohort Profile: The Care Trajectories—Enriched Data (TorSaDE) cohort

Alain Vanasse,^{1,2,3,*†} Yann M Chiu,^{1,2} Josiane Courteau,²
Marc Dorais,⁴ Gillian Bartlett,⁵ Kristina Zawaly ⁵ and Mike Benigeri^{3,6}



II – Grande utilisation des SU

Un peu de code  utile...

```
summary(data)
```

```
str(data)
```

```
dim(data)
```

```
head(data)
```

```
data$newVar <- ifelse(age > 40, 1, 0)
```

```
as.factor(var)
```

```
as.numeric(var)
```

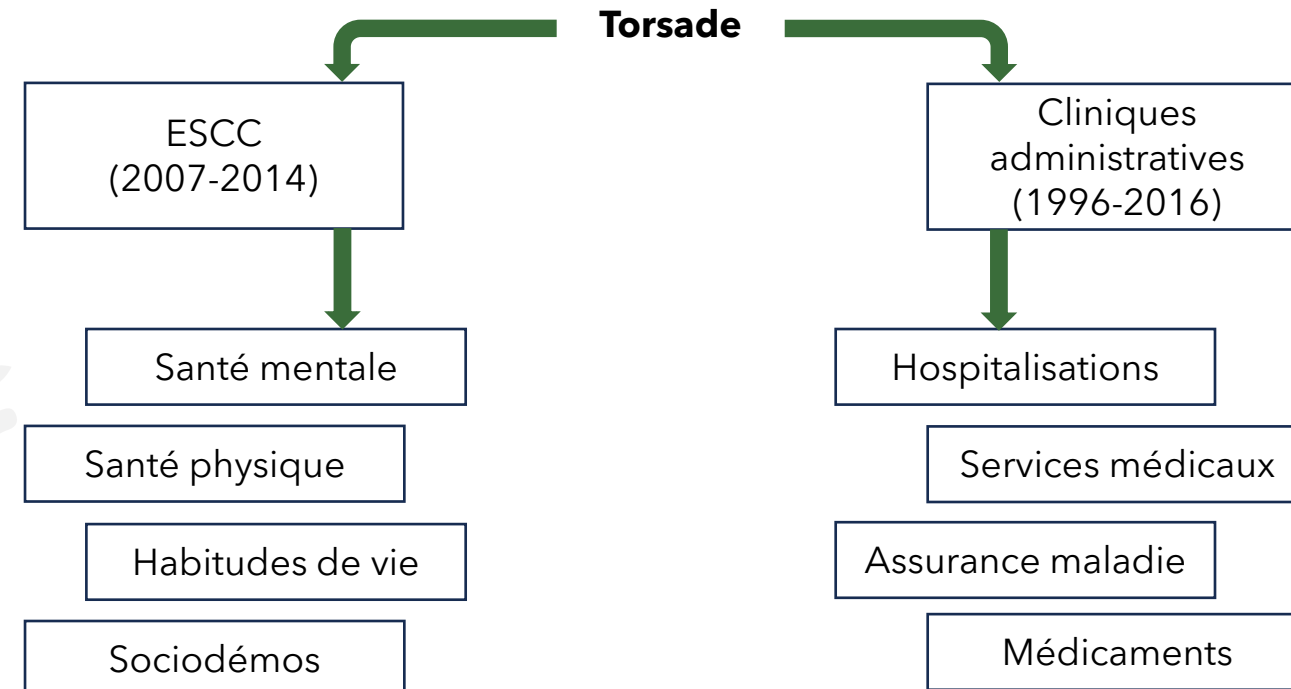
```
??function
```



Cohort Profile

Cohort Profile: The Care Trajectories—Enriched Data (TorSaDE) cohort

Alain Vanasse,^{1,2,3*†} Yann M Chiu,^{1,2} Josiane Courteau,²
Marc Dorais,⁴ Gillian Bartlett,⁵ Kristina Zawaly ⁵ and Mike Benigeri^{3,6}

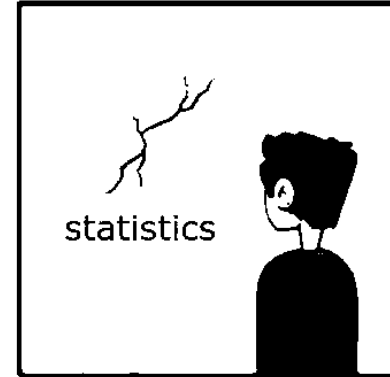


II – Grande utilisation des SU

Vocabulaire: apprentissage automatique (*machine learning*)

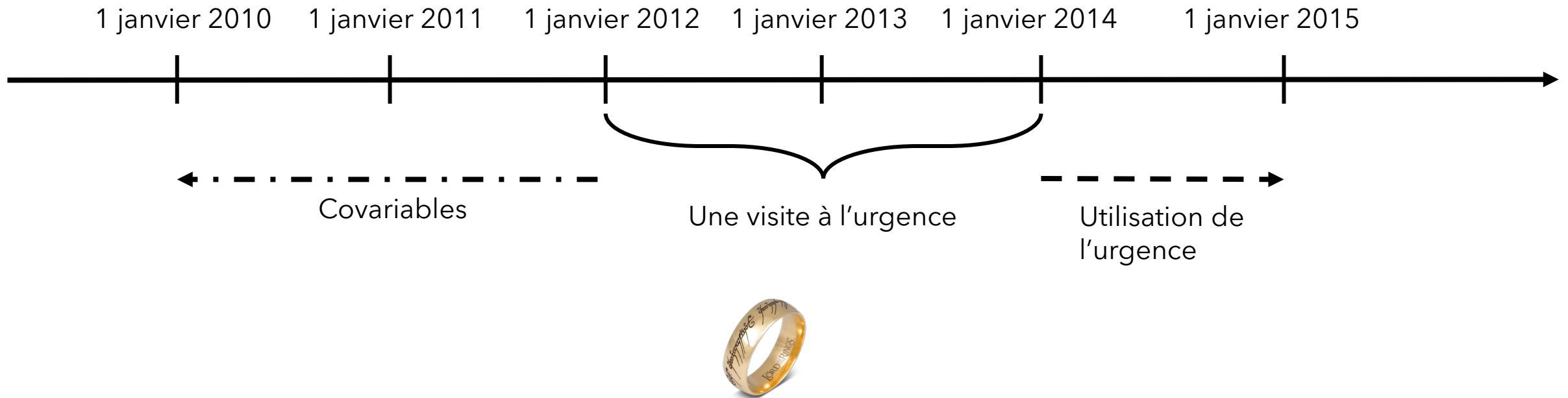
- Apprentissage supervisé (prédiction d'issues de santé)
- Apprentissage non supervisé (profils d'individus)
- Apprentissage par renforcement (un chat est un chat)

Beaucoup de concepts se recoupent entre la biostatistique et l'apprentissage automatique



Adapté de <https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3>

II – Grande utilisation des SU



II – Grande utilisation des SU

Analyse de classes latentes: modèle probabiliste qui identifie des sous-groupes de population

Les individus au sein d'un même groupe sont similaires, tout en étant différents à ceux des autres groupes

1) Faible morbidité , 2) Comorbidité élevée, 3) Santé mentale ou abus de substances, 4) Blessures ou douleurs

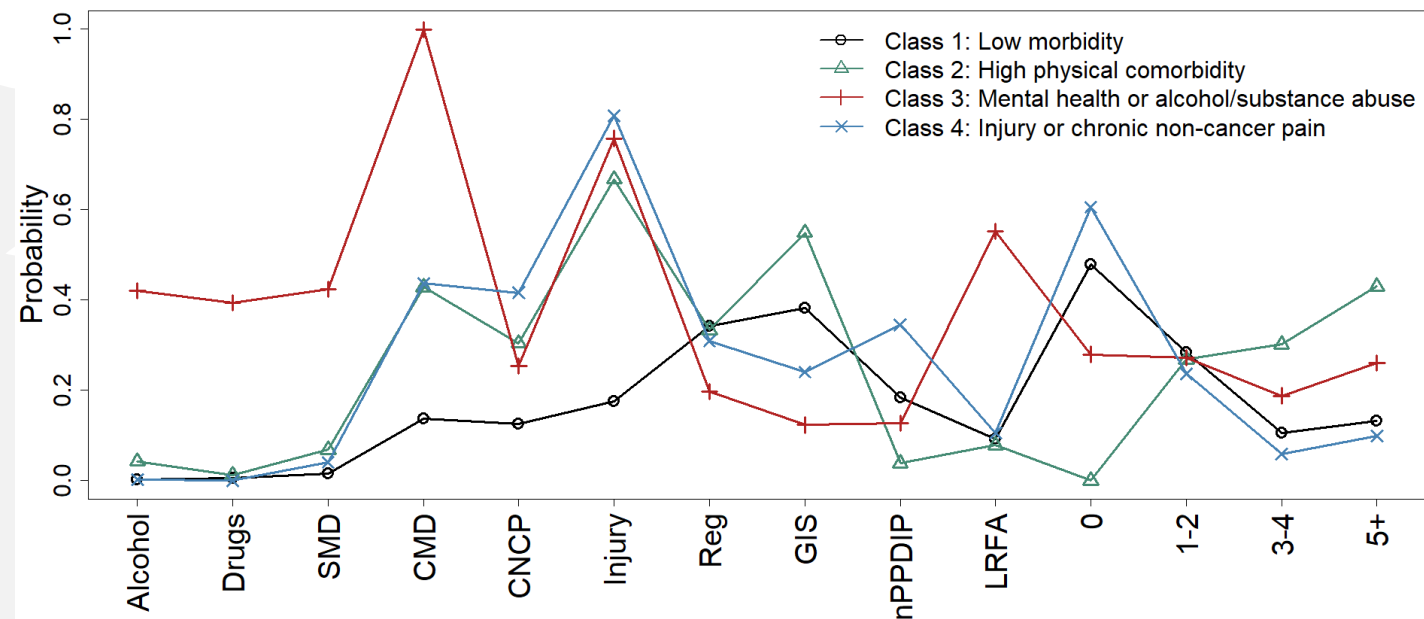
1. Chiu, Y. M., et al. (2022). Profiles of frequent emergency department users with chronic conditions: a latent class analysis. *BMJ open*, 12(9), e055297.

Open access

Original research

BMJ Open Profiles of frequent emergency department users with chronic conditions: a latent class analysis

Yohann Moanahere Chiu^{1,2}, Isabelle Dufour^{3,4}, Josiane Courteau²,
Alain Vanasse^{1,2}, Maud-Christine Chouinard⁵, Marie-France Dubois⁶,
Nicole Dubuc^{3,7}, Nicolas Elazhary¹, Catherine Hudon^{1,2}



SMD: Serious mental disorders, CMD: Common mental disorders, CNCP: Chronic non-cancer pain. Reg, GIS, nPPDIP, and LRFA refer to PPDIP status while 0, 1-2, 3-4, and 5+ refer to the comorbidity index.

II – Grande utilisation des SU

Régression logistique pour identifier les facteurs cliniques les plus importants

Jumelage avec des questionnaires auto-rapportés (facteurs culturels et environnementaux)

Intégration des résultats dans un **outil de dépistage** des patients aux besoins complexes

1. Hudon, C., Bisson, M., Dubois, M. F., Chiu, Y., Chouinard, M. C., Dubuc, N., ... & Vanasse, A. (2021). CONECT-6: a case-finding tool to identify patients with complex health needs. *BMC health services research*, 21, 1-9.

[Home](#) > [BMC Health Services Research](#) > [Article](#)

CONECT-6: a case-finding tool to identify patients with complex health needs

Research article | [Open access](#) | Published: 17 February 2021

Volume 21, article number 157, (2021) [Cite this article](#)

Questions	Answers	
	Yes	No
1. In general, would you say your health is fair or even poor?		
2. Do you have pain or discomfort preventing most of your activities?		
3. In the past 12 months, do you consider your health needs were met less than half of the time?		
4. Do your interactions with the health system and health professionals ever make you feel like you have complex health problems?		
5. Is your household income from all sources before taxes and other deductions less than \$20,000?		
6. In the past 12 months, have you rarely or even never received support from friends or relatives when you needed it?		
Number of yes and no answers	— —	— —

- I. Les banques de données médico-administratives pour la recherche
 - Avantages et inconvénients
 - SISMACQ
- II. Grande utilisation des services d'urgence
 - Profils des grands utilisateurs
 - Construction d'un outil de dépistage
- III. Surveillance de la polypharmacie
 - Fouille de données
 - Prédiction
 - Transparence
- IV. Conclusions et réflexions
 - On n'a pas parlé de...
 - Statistique et IA pour la recherche en santé

III – Surveillance de la polypharmacie

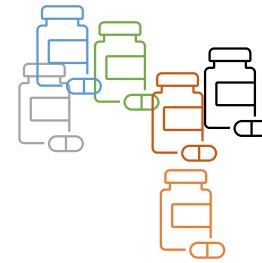
Polypharmacie associée à des **issues de santé négatives** (décès, fragilité, CHSLD, hospitalisation)¹

Surveillance accrue chez les populations vulnérables²

Difficile à départager étant donné le nombre impressionnant de combinaisons possibles

➡ **méthodes IA?**

1. Pazan, F., & Wehling, M. (2021). Polypharmacy in older adults: a narrative review of definitions, epidemiology and consequences. *European geriatric medicine*, 12, 443-452.
2. Gosselin, M., Talbot, D., Simard, M., Chiu, Y. M., ... & Sirois, C. (2023). Classifying Polypharmacy According to Pharmacotherapeutic and Clinical Risks in Older Adults: A Latent Class Analysis in Quebec, Canada. *Drugs & Aging*, 40(6), 573-583.



3500 dénominations communes



{Rx1; Rx2}



$6 \cdot 10^6$ possibilités

{Rx1; ...; Rx3}



$7 \cdot 10^9$ possibilités

{Rx1; ...; Rx4}



$6 \cdot 10^{12}$ possibilités

{Rx1; ...; Rx5}



$4 \cdot 10^{15}$ possibilités

1 combinaison/seconde ~ 130 millions d'années



III – Surveillance de la polypharmacie

L'IA promet des résultats plus précis et plus rapides en épidémiologie

TRÈS populaire en médecine personnalisée... et en surveillance?

Deux méthodes de fouille de données pour déterminer les **séquences** et les **combinaisons fréquentes**

1. Bukhtiyarova, O., Abderrazak, A., Chiu, Y., Sparano, S., Simard, M., & Sirois, C. (2022). Major areas of interest of artificial intelligence research applied to health care administrative data: a scoping review. *Frontiers in Pharmacology*, 13, 944516.

Georg Thieme Verlag KG Stuttgart

Artificial Intelligence in Public Health and Epidemiology

Rodolphe Thiébaud, Frantz Thiessard, Section Editors for the IMIA Yearbook Section on Public Health and Epidemiology Informatics

[> Author Affiliations](#)

Challenges and opportunities for public health made possible by advances in natural language processing

[Oliver Baclic](#),^{1,*} [Matthew Tunis](#),¹ [Kelsey Young](#),¹ [Coraline Doan](#),² [Howard Swerdfeger](#),² and [Justin Schonfeld](#)^{3,*}

► Author information ► Copyright and License information [Disclaimer](#)

¹ Faculty of Pharmacy, Université Laval, Québec, QC, Canada

² Quebec National

³ Faculty of Medicine

Home » American Journal of Public Health (AJPH) » January 2021

Artificial Intelligence, Intersectionality, and the Future of Public Health

Greta R. Bauer PhD, MPH, and Daniel J. Lizotte PhD

[+] Author affiliations, information, and correspondence details

Accepted: October 09, 2020 Published Online: December 16, 2020

Introduction
intelligence
care admini

Methods: T
digital libr

de

sc

in

Re

be

pa

fo

ha

m

M

Review Article | [Open Access](#) | [Published: 26 February 2021](#)

Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies

[Dinesh Visva Gunasekeran](#), [Rachel Marjorie Wei Wen Tseng](#), [Yih-Chung Tham](#) & [Tien Yin Wong](#) [✉](#)

[npj Digital Medicine](#) 4, Article number: 40 (2021) | [Cite this article](#)

11k Accesses | 48 Citations | 47 Altmetric | [Metrics](#)

Conclusions: The scoping review revealed the potential of AI application to health-related studies. However, several areas of interest in pharmacoepidemiology are sparsely reported, and the lack of details in studies related to pharmacotherapy suggests that AI could be used more optimally in pharmacoepidemiologic research.

Finally, the performance of digital health technology for operational applications related to **population surveillance** [...] have not been adequately evaluated"

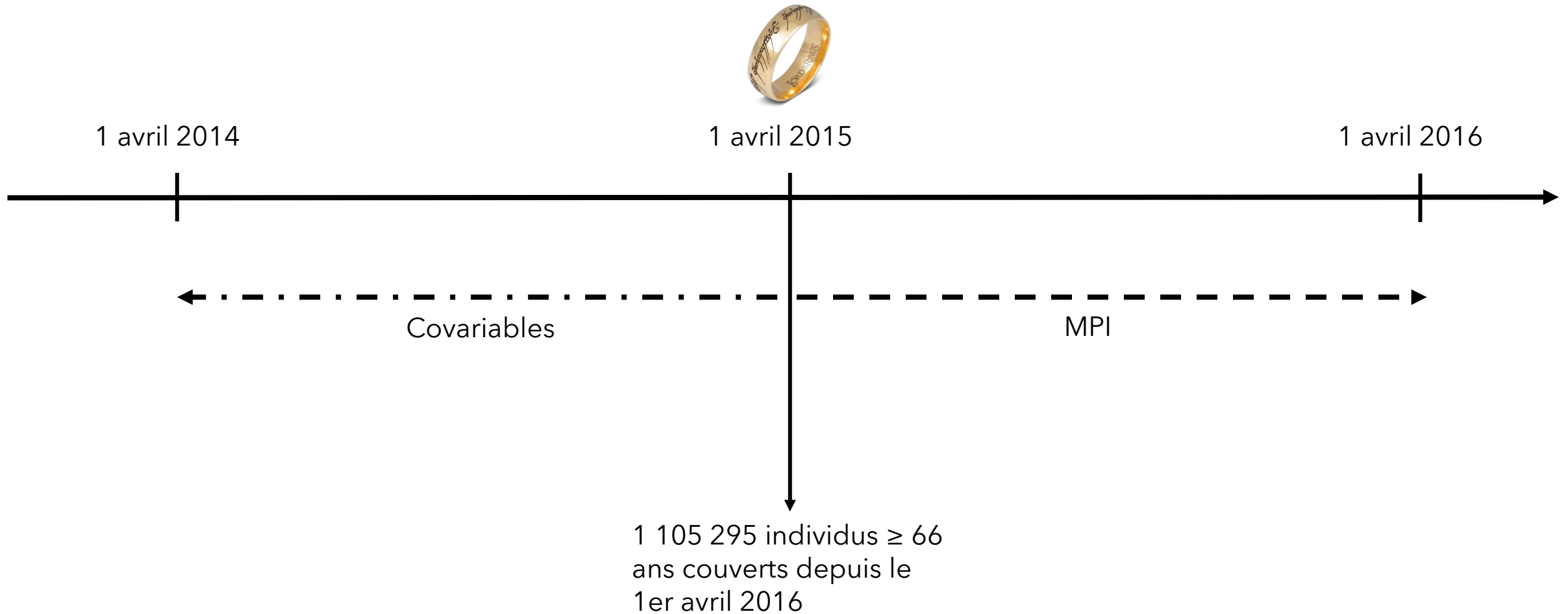
arch

view

h outcomes data. We
extract health areas of

, the most common
ical data, clinical
and analysis-friendly
des (15%). Less attention
nent learning (1%). The
upport vector machines.
nophen, and heparin.

III – Surveillance de la polypharmacie

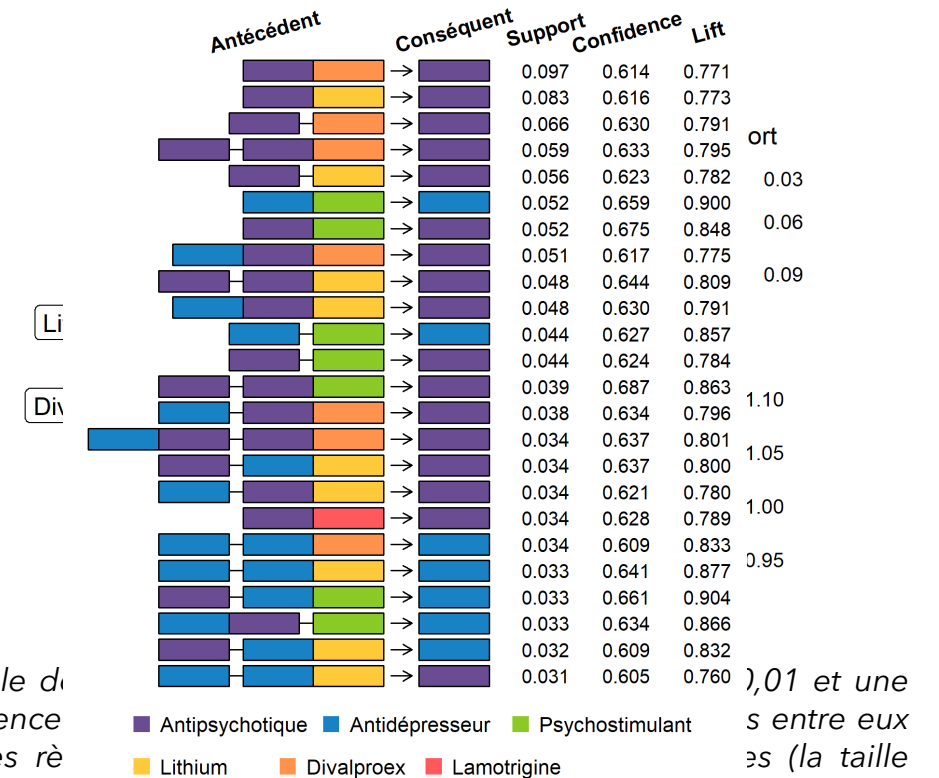


III – Polypharmacie et fouille de données

Combinaison ou séquence fréquente se mesure avec le **support**

Prescrire du lithium augmente de 10% la probabilité de prescrire un antipsychotique (**lift**)

Antipsychotique est prescrit dans 61% des cas lorsque la combinaison antipsychotique + divalproex est d'abord prescrite (**confidence**)



Exemple de séquences avec un seuil de 0,03 et une confiance minimale de 0,6 pour l'ARM. Pour chaque séquence (ligne), le médicament à droite de la flèche est prescrit lorsque la ou les séquences (séparées par des traits horizontaux) ont d'abord été prescrites (à gauche de la flèche).

III – Polypharmacie et prédiction

Régression logistique: modèle de régression pour variable binaire

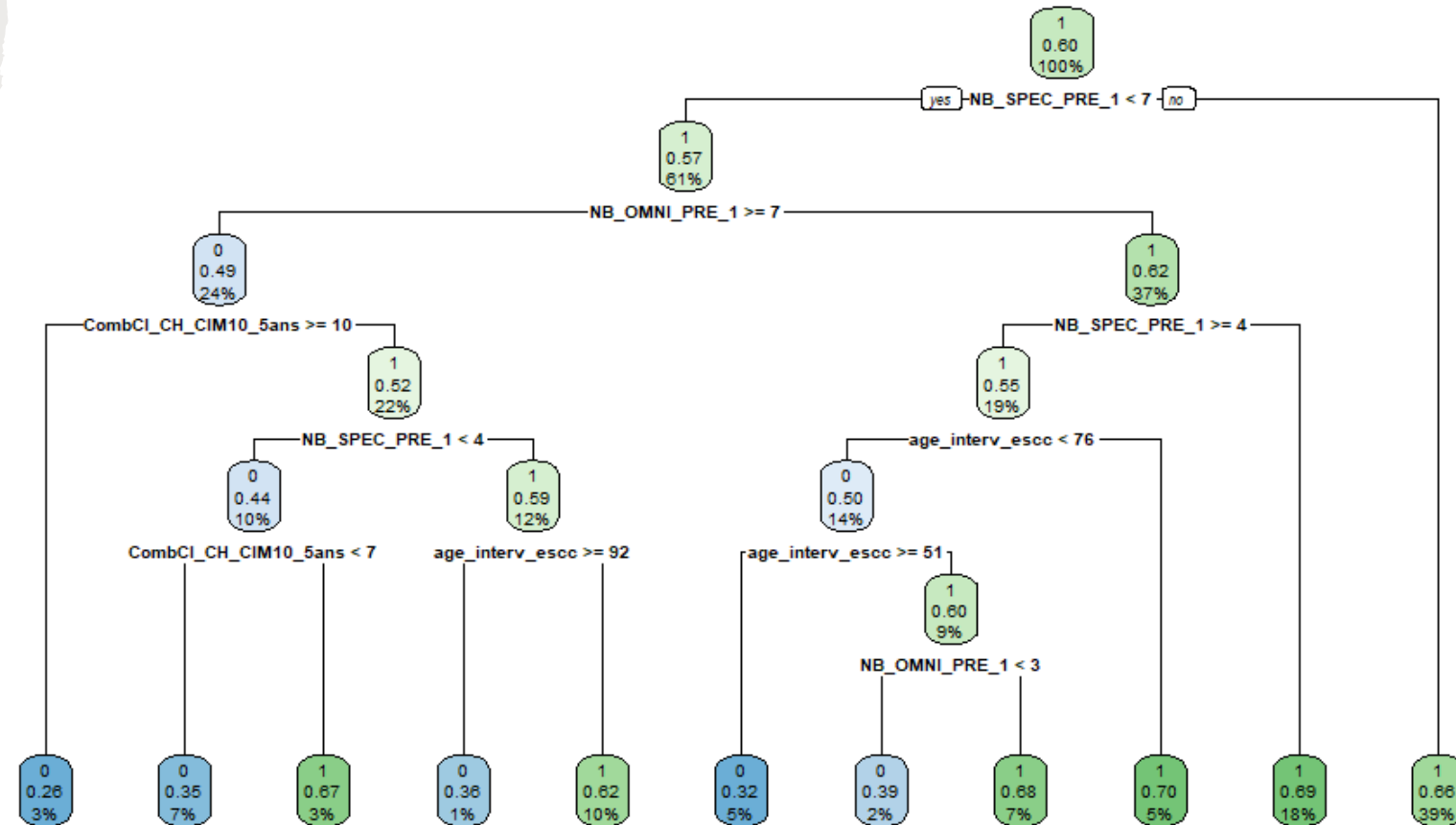
Gradient boosting machines: construit un ensemble de souches

Naïve Bayes: application du théorème de Bayes

Forêts aléatoires: arbres de décisions successifs

Réseaux de neurones: « imitation » des neurones biologiques

Métriques: sensibilité, spécificité, VPP, VPN, Brier (prédictions-observations)²



III – Polypharmacie et prédiction

Régression logistique: modèle de régression pour variable binaire

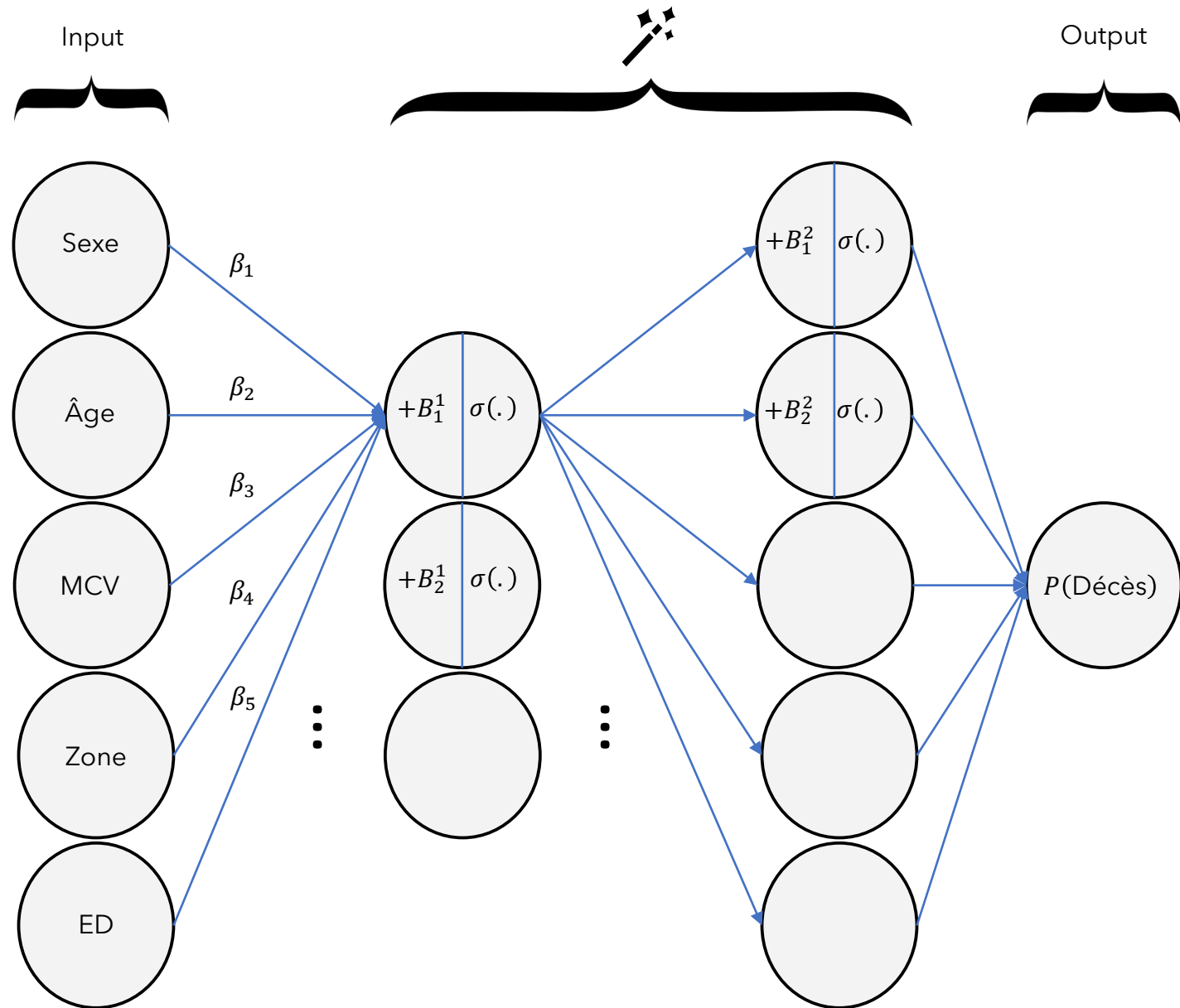
Gradient boosting machines: construit un ensemble de souches

Naïve Bayes: application du théorème de Bayes

Forêts aléatoires: arbres de décisions successifs

Réseaux de neurones: « imitation » des neurones biologiques

Métriques: sensibilité, spécificité, VPP, VPN, Brier (prédictions-observations)²



III – Polypharmacie et prédiction

Régression logistique: modèle de régression pour variable binaire

Gradient boosting machines: construit un ensemble de souches

Naïve Bayes: application du théorème de Bayes





Forêts aléatoires: arbres de décisions successifs

Réseaux de neurones: « imitation » des neurones biologiques

Métriques: sensibilité, spécificité, VPP, VPN, Brier (prédictions-observations)²

Observations

Prédictions

		
	VP	FN
	FP	VN

$$\text{Sen} = \frac{VP}{VP+FN}$$

$$\text{Spé} = \frac{VN}{FP+VN}$$

$$\text{VPP} = \frac{VP}{VP+FP}$$

$$\text{VPN} = \frac{VN}{FN+VN}$$

III – Polypharmacie et prédiction

1 105 295 aînés dans la province du Québec (2015-2016)

Prédiction de l'utilisation de **médicaments potentiellement inappropriés** (liste de Beers)

Ajustés pour l'âge, le sexe, la défavorisation sociale/matérielle, la zone de résidence, les maladies chroniques, le nombre d'hospitalisations



Modèle	AUC 95% CI	Sen	Spé	VPP	VPN	Brier
RL	0.62 0.62 - 0.62	0.49	0.73	0.62	0.61	0.38
GBM	0.62 0.62 - 0.62	0.49	0.73	0.63	0.61	0.38
NB	0.61 0.61 - 0.61	0.38	0.80	0.63	0.59	0.40
RN	0.62 0.62 - 0.62	0.52	0.70	0.62	0.62	0.38
FA	0.62 0.62 - 0.62	0.51	0.72	0.62	0.62	0.38

III – Polypharmacie et prédiction

Qu'en est-il de l'**importance** des variables dans la prédiction?

Différentes mesures pour l'importance selon le modèle

Importance **globale** semble concordante~ish

1. Chiu, YM., Sirois, C., Simard, M., Gagnon, ME. & Talbot, D. (2024). Traditional methods hold their ground against machine learning in predicting potentially inappropriate medication use in older adults. *Value in Health*, revisions required.

Variable	RL	FA	GBM	NB	RN
Troubles mentaux	1	1	1	7	15
Sexe	2	2	2	2	2
Diabètes	3	4	5	6	12
Hypertension	4	5	3	1	7
MPOC	5	3	4	4	10
Hospitalisation	6	7	6	9	6
MCV	7	9	7	5	8
Alzheimer	8	6	10	14	14
Asthme	9	8	9	10	11
Zone résidentielle	10	10	12	15	3
Âge	11	11	8	3	1
Ostéoporose	12	12	11	8	13
Défavorisation sociale	13	14	14	11	5
Insuffisance cardiaque	14	13	13	12	9
Défavorisation matérielle	15	15	15	13	4

+vert = +important

III – Polypharmacie et prédiction

Pas de vainqueur clair pour les performances...

On n'a pas pris en compte les facteurs liés à l'utilisation des médicaments et le style de vie

Dans la boîte noire... d'autres boîtes noires

Comment expliquer la performance des RN?



Variable	RL	FA	GBM	NB	RN
Troubles mentaux	1	1	1	7	15

III – Polypharmacie et prédiction

À vous!

- Explorer les forêts aléatoires (package randomForest)
- Explorer les réseaux de neurones (package nnet)
- Explorer l'importance des variables (package caret)

III – Polypharmacie et prédiction

Un peu de code  utile...

```
model <- randomForest(formula,  
data, ntrees)  
model <- nnet(formula, data, size)  
  
varImp(model)  
table(observations, predictions)
```

Exemple de formule:

Death ~ Age + Sex + MCV

III – Polypharmacie et explicabilité

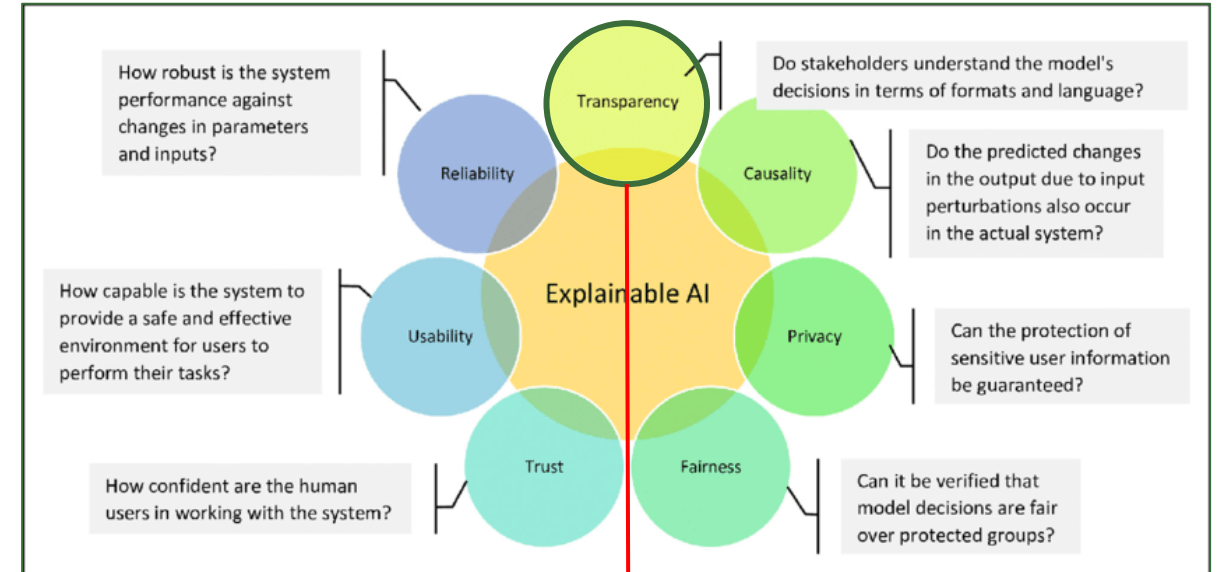
IA \approx boîte noire

Transparence: deux caractéristiques différentes en XAI

Être capable d'expliquer une prédiction de l'algorithme aide à la **généralisabilité/confiance**

Rendre la boîte noire... plus grise?

Source: Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, 19(2), 133-144.



Explicabilité

\neq

Interprétabilité

Expliquer une sortie de l'algorithme

Suivre pas à pas les étapes de l'algorithme

III – Polypharmacie et explicabilité

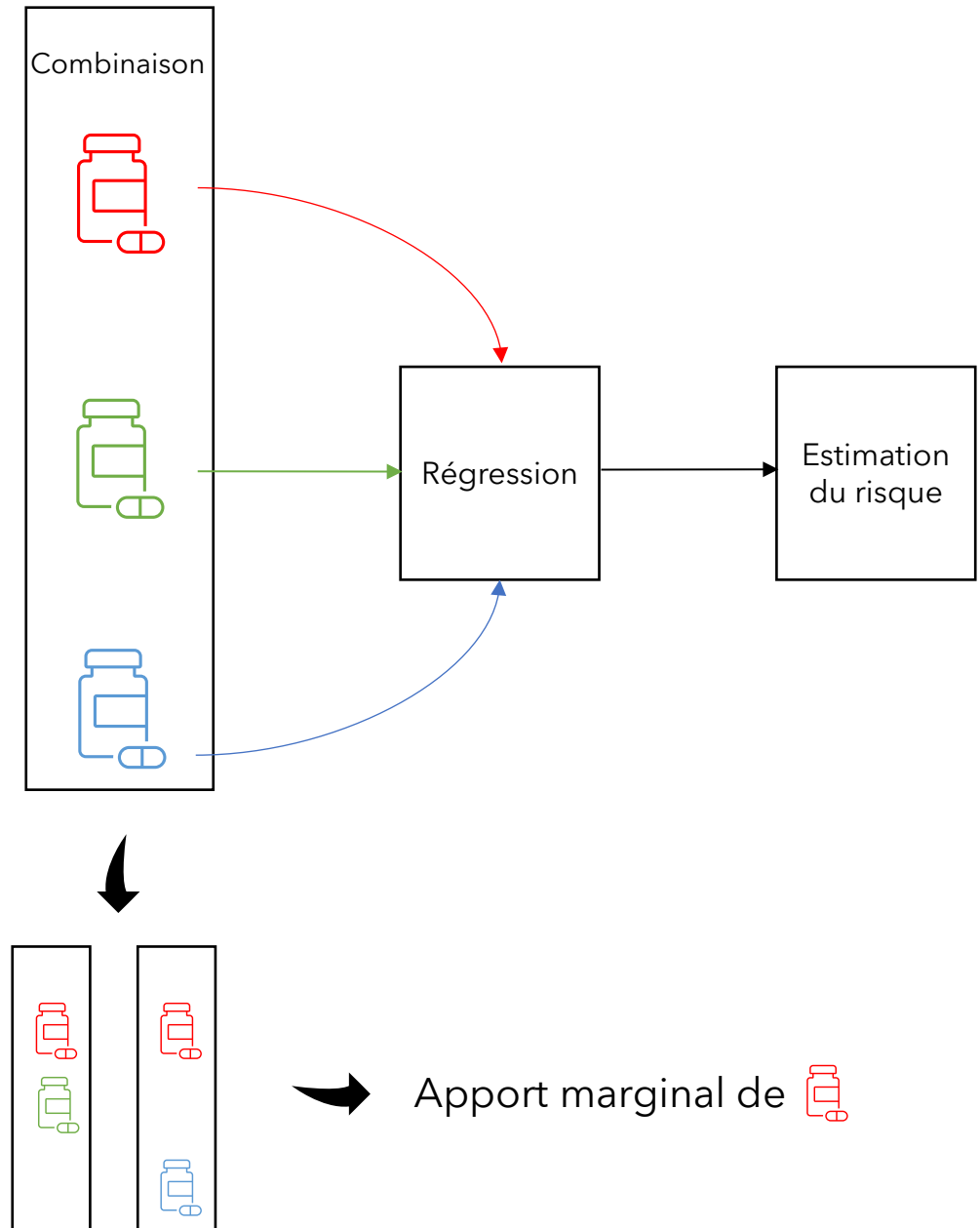
Valeurs SHAP¹

Théorie des jeux: dans une équipe de N joueurs possibles, quel est l'apport du joueur i dans le gain total v ?

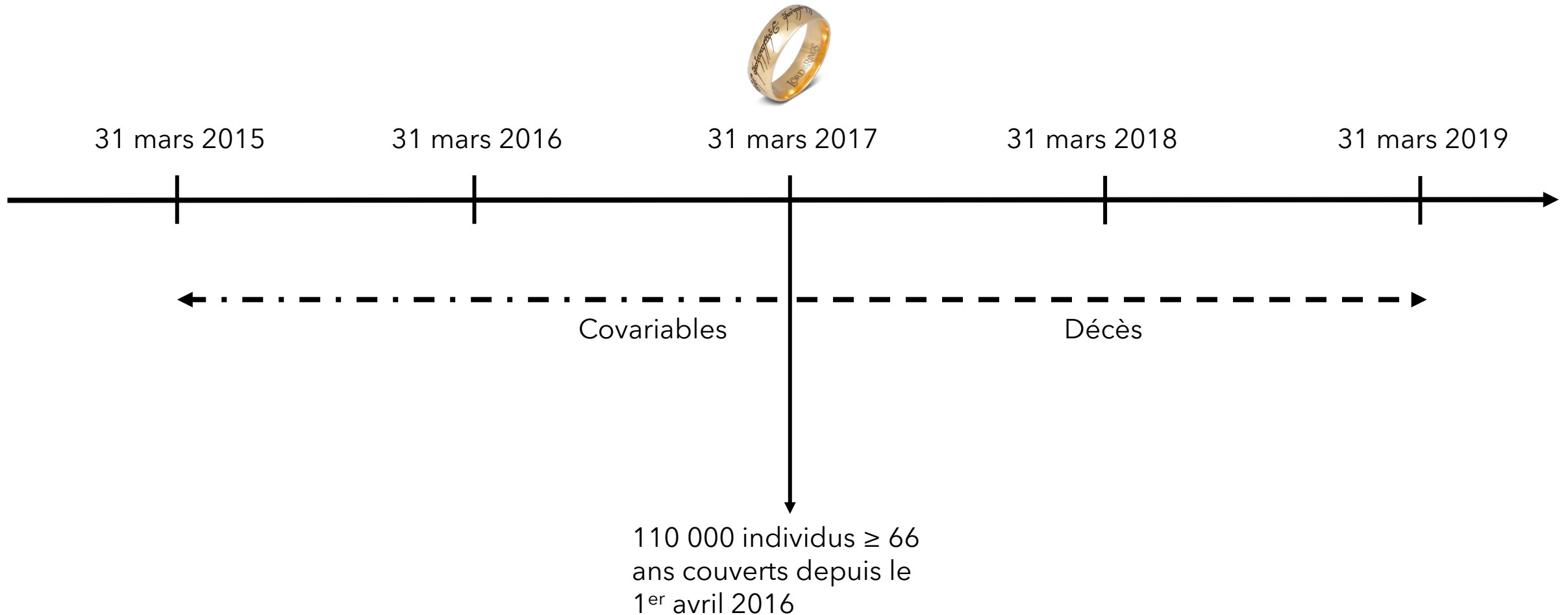
Il faut prendre en compte:

- Le « poids » de chaque joueur
- Les interactions entre les joueurs
- L'ordre d'arrivée des joueurs

1. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.



III – Polypharmacie et explicabilité

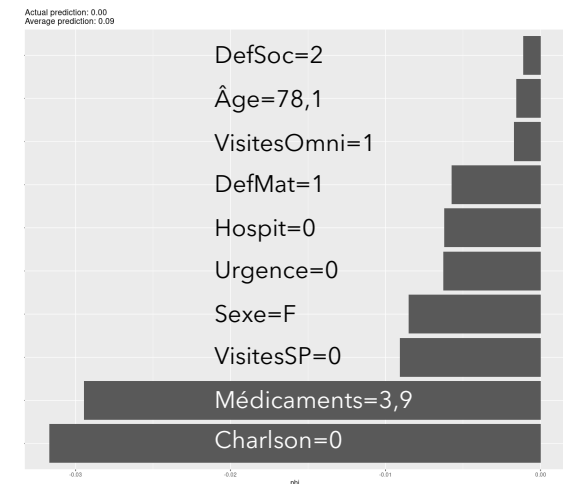
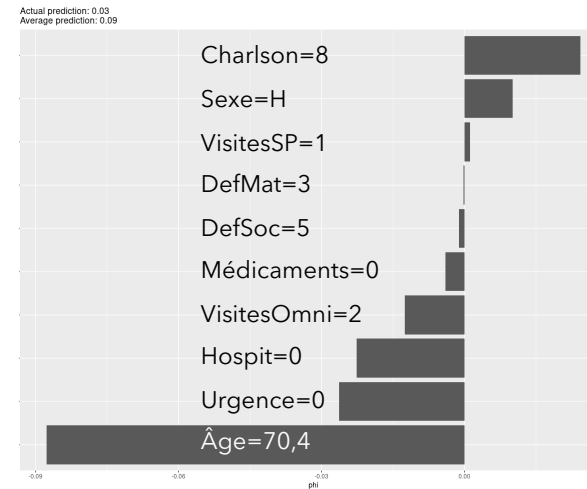
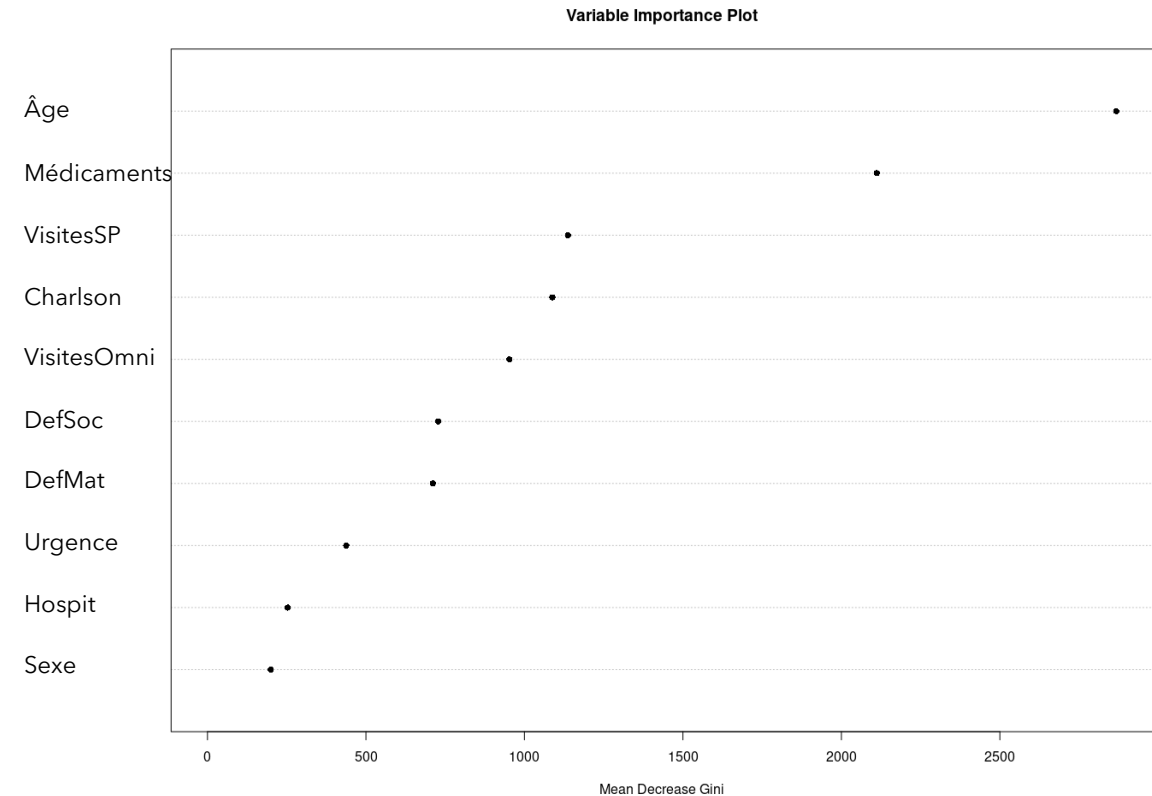


III – Polypharmacie et explicabilité

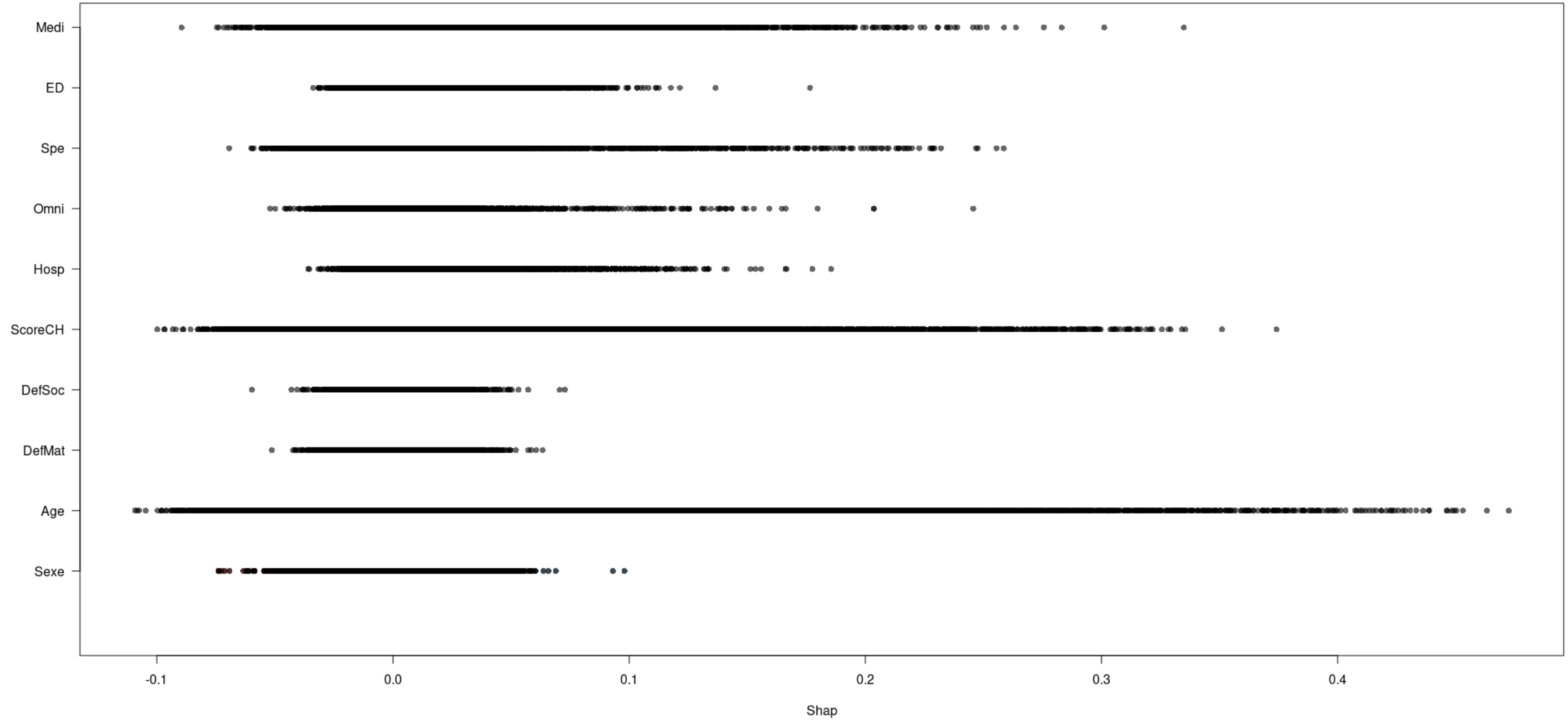
Tous les modèles AI/ML/stats mesurent l'importance **globale**

SHAP permet une inspection **locale** de l'importance, c'est-à-dire pour un individu

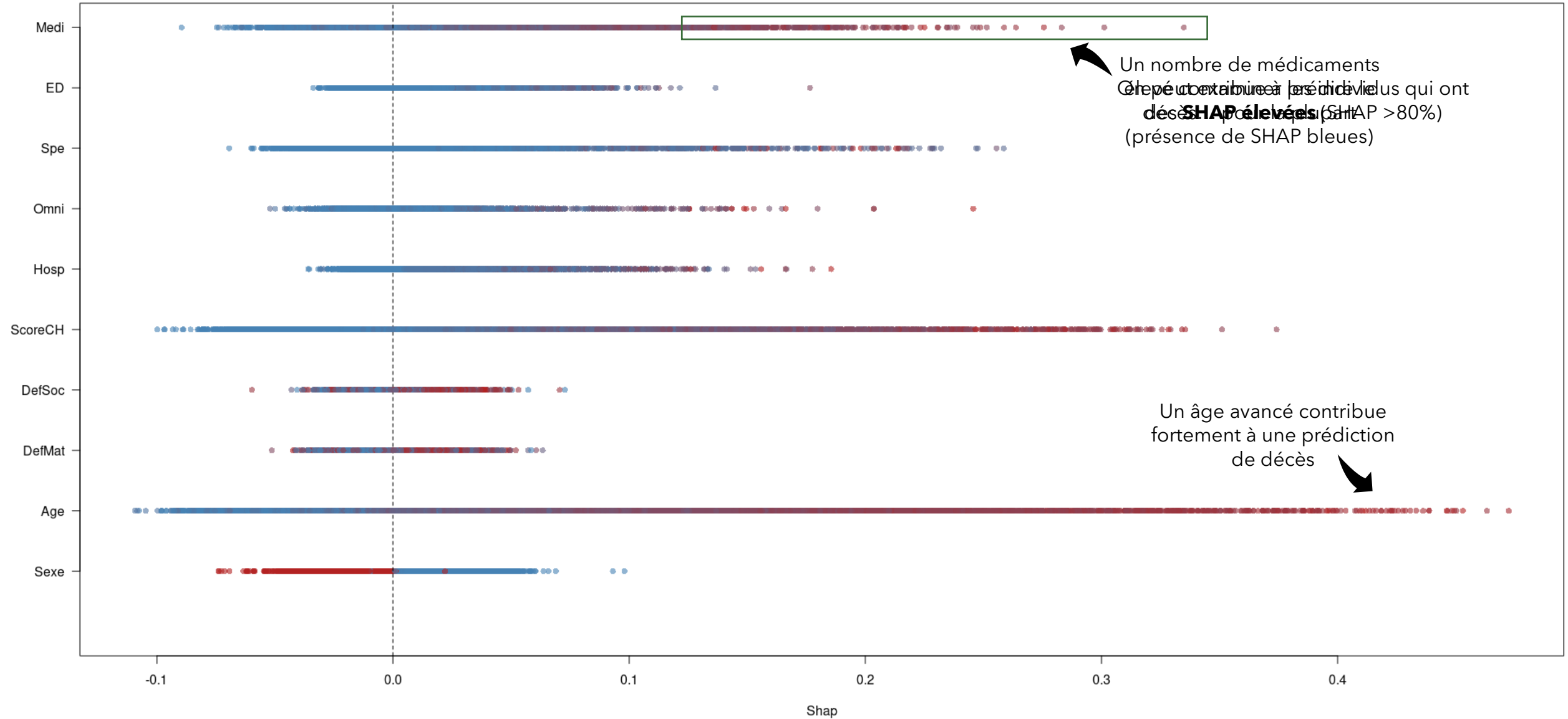
Quelles valeurs de quelles variables ont eu un impact sur la prédiction?



Valeur SHAP élevée « pousse » le
modèle à prédire le décès



Rouge: valeurs élevées de la variable
Bleu: valeurs basses de la variable



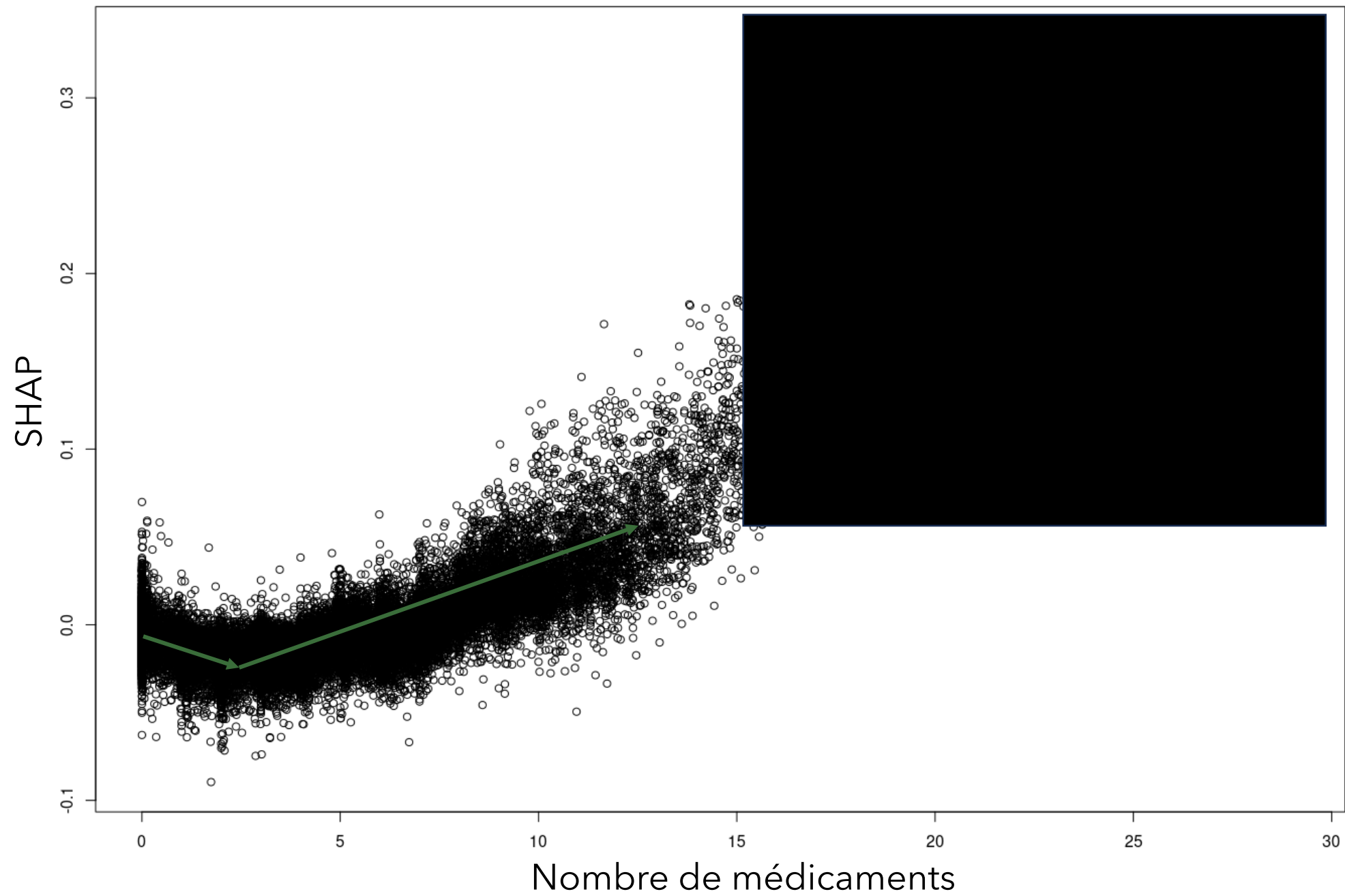
III – Polypharmacie et explicabilité

La population SHAP>80% a des **caractéristiques similaires** avec une population >8 médicaments

SHAP>80% est plus intéressante puisqu'il y a des individus qui ont 0, 1, 2... médicaments

SHAP permet également d'observer la forme de la relation du pouvoir prédictif

Variable (moyenne)	Population générale	SHAP > 80%
Age	75,6	78,6
Femme %	55,6	55,5
Score de Charlson	1,6	3,3
Nombre d'hospitalisations	0,1	0,3
Nombre de visites à l'urgence	0,5	0,8
Nombre de visites médecin spécialiste	4,3	6,7
Nombre de médicaments - médiane (Q1-Q3)	4,6 (2,1-7,3)	10,1 (8,5-12,3)



III – Polypharmacie et explicabilité

SHAP

- Permet d'examiner l'hétérogénéité dans la prédiction
- Long... très long: 2 jours de calcul pour ~30 000 individus

Il existe d'autres méthodes d'explicabilité des modèles IA

Chacune présente des défis et avantages distincts

Aucune n'est une fin en soi, mais toute méthode d'explicabilité est utile pour la prédiction


scientific reports

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [scientific reports](#) > [articles](#) > article

Article | [Open access](#) | [Published: 04 April 2023](#)

Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME

[Mariam Laatifi](#), [Samira Douzi](#) , [Hind Ezzine](#), [Chadia El Asry](#), [Abdellah Naya](#), [Abdelaziz Bouklouze](#), [Younes Zaid](#) & [Mariam Naciri](#)

[Scientific Reports](#) **13**, Article number: 5481 (2023) | [Cite this article](#)

1875 Accesses | 4 Citations | 5 Altmetric | [Metrics](#)

- I. Les banques de données médico-administratives pour la recherche
 - Avantages et inconvénients
 - SISMACQ
- II. Grande utilisation des services d'urgence
 - Profils des grands utilisateurs
 - Construction d'un outil de dépistage
- III. Surveillance de la polypharmacie
 - Fouille de données
 - Prédiction
 - Transparence
- IV. Conclusions et réflexions
 - On n'a pas parlé de...
 - Statistique et IA pour la recherche en santé

IV – Conclusions et réflexions

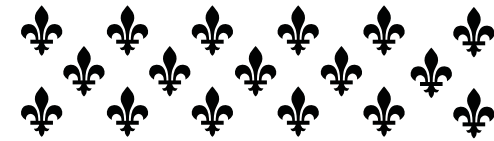


On n'a pas parlé...

- Des enjeux de protection de la vie privée, d'éthique de la recherche
- D'analyse fédérée qui permet de ne pas déplacer les données

En résumé...

- Les banques de données clinico-administratives offrent des opportunités uniques pour la recherche populationnelle
- Il existe beaucoup d'outils d'analyse statistique (BEAUCOUP)
- L'IA gagne en popularité



ASSEMBLÉE NATIONALE DU QUÉBEC

DEUXIÈME SESSION

QUARANTE-DEUXIÈME LÉGISLATURE

Projet de loi n° 19

**Loi sur les renseignements de santé
et de services sociaux et modifiant
diverses dispositions législatives**

Présentation

Présenté par
M. Christian Dubé
Ministre de la Santé et des Services sociaux

Éditeur officiel du Québec
2021

« L'ampleur du SISMACQ, qui contient plusieurs millions d'informations dans une **vingtaine de bases de données** structurées selon un modèle relationnel complexe, engendre plusieurs défis pour son utilisation. Le **temps d'exécution et l'espace mémoire** disponible représentent aussi des défis d'utilisation »

INSPQ INSTITUT NATIONAL
DE SANTÉ PUBLIQUE
DU QUÉBEC



Cadre de qualité des données du Système intégré
de surveillance des maladies chroniques du Québec

RAPPORT MÉTHODOLOGIQUE

IV – Conclusions et réflexions



On n'a pas parlé...

- Des enjeux de protection de la vie privée, d'éthique de la recherche
- D'analyse fédérée qui permet de ne pas déplacer les données

En résumé...

- Les banques de données clinico-administratives offrent des opportunités uniques pour la recherche populationnelle
- Il existe beaucoup d'outils d'analyse statistique (BEAUCOUP)
- L'IA gagne en popularité

Renseignements fiables à partir de processus d'apprentissage automatique responsables

RESPECT DES PERSONNES

- Valeur pour les Canadiens
- Prévention des préjudices
- Équité
- Imputabilité

RESPECT DES DONNÉES

- Vie privée
- Sécurité
- Confidentialité



APPLICATION RIGOUREUSE

- Transparence
- Reproductibilité du processus et des résultats

MÉTHODES ÉPROUVÉES

- Qualité des données d'apprentissage
- Inférence valide
- Modélisation rigoureuse
- Explicabilité

Évaluation au moyen de l'autoévaluation et de l'examen par les pairs, à l'aide d'une liste de vérification et de la production d'un rapport ou d'un tableau de bord

<https://www.statcan.gc.ca/fr/science-donnees/reseau/apprentissage-automatique>

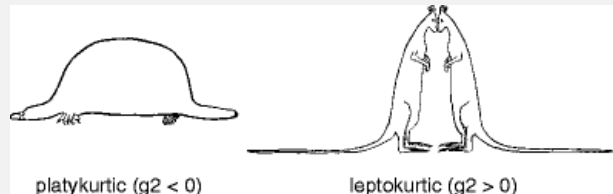
IV – Conclusions et réflexions



Les méthodes IA nécessitent de la réflexion en amont de l'application

Par ex. : Vocabulaire différent

Par ex. bis: En statistique, on suppose des modèles (omniprésence de la loi normale); en IA/ML, on laisse « parler les données »






Wright, D.B., Herrington, J.A. Problematic standard errors and confidence intervals for skewness and kurtosis. *Behav Res* 43, 8-17 (2011).

Biostatistique Épidémiologie	AI/ML
Estimation	Apprentissage
Régression	Apprentissage supervisé
Sensibilité/Valeur prédictive positive	Rappel/Précision
Variable dépendante	Cible
Paramètres	Poids
Variable indépendante	Caractéristique (<i>feature</i>)
Observation	Instance
Aire sous la courbe ROC	
Surajustement	
Modèle	

IV – Conclusions et réflexions

L'application de IA/ML ne devrait pas être un but ultime 


On augmente la boîte (noire) à outils   




Pas de gain (majeur) de performances avec des données traitées et structurées

Surtout: besoin de s'accorder sur des objectifs communs!

Biostatistique Épidémiologie	AI/ML
Estimation	Apprentissage
Régression	Apprentissage supervisé
...	
Expliquer associations	Estimer des paramètres
Santé des populations	Algorithmes optimaux
...	
But commun!	

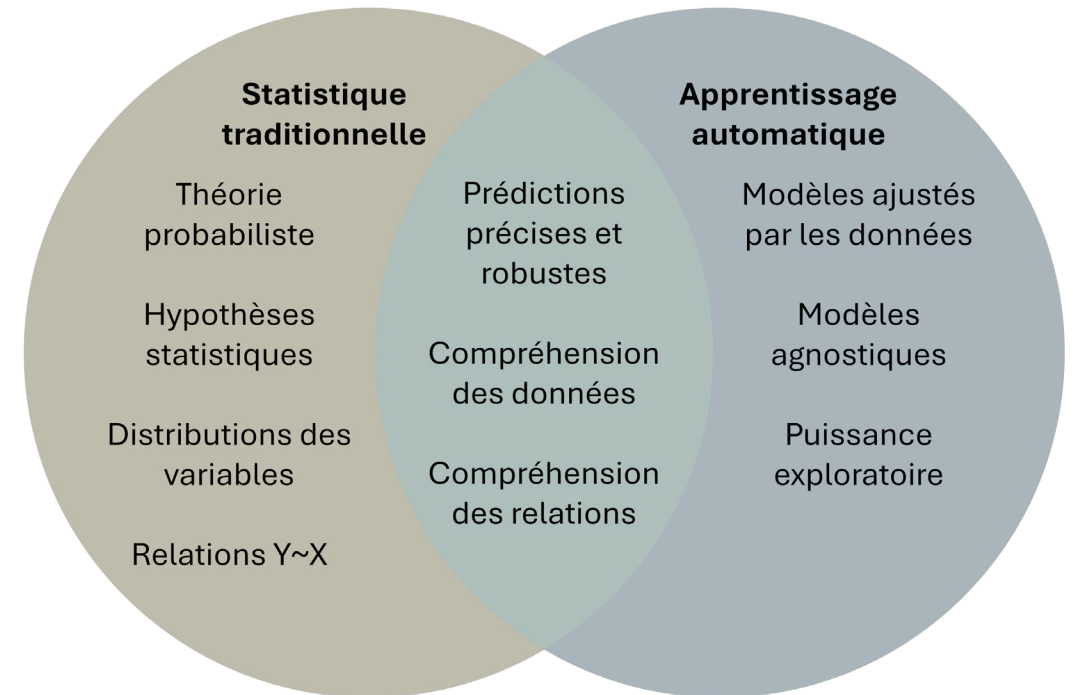
IV – Conclusions et réflexions

L'application de IA/ML ne devrait pas être un but ultime 

On augmente la boîte (noire) à outils   

Pas de gain (majeur) de performances avec des données traitées et structurées

Surtout: besoin de s'accorder sur des objectifs communs!



Merci! Avez-vous des questions pour moi?

Traductions libres

« Tous les modèles sont faux, certains sont utiles » (George Box)

« Il n'est pas très utile de dire que tous les modèles sont faux [...]. Ne pas tomber en amour d'un seul modèle pour exclure les autres » (Peter McCullagh et John Nelder)

« Les statistiques sont la grammaire des sciences » (Karl Pearson)