Traitement de la langue naturelle pour une réponse rapide aux maladies émergentes: COVID-19

Antoine Neuraz, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, Kevin Bretonnel Cohen, Anita Burgun, Nicolas Garcelon, Bastien Rance,

AP-HP/Universities/INSERM COVID-19 Research Collaboration

4 février 2021

Necker ENFANTS MALADES HÔPITAL UNIVERSITAIRE

UNIVERSITÉ PARIS DESCARTES

Université de Paris

Inserm

cnrs

Limsi

Qu’est-ce qu’un dossier patient informatisé (DPI) ?

Usages

Documentation

Recherche d’information

Communication

- Transmissions
- Prescriptions
- Demandes d’examen

Pour la Recherche

Données

Formulaires

- demandes d’examens, …

Résultats d’examens

- Biologie, Imagerie, …

Prescriptions

- médicaments, actes, …

Textes cliniques

- compte-rendu, lettres, …

Codage

- Actes, Diagnostics, …

…

Médecine basée sur les preuves ➔ données structurées

Faciles à requêter, stocker, analyser

MAIS

Formulaires (fastidieux ?)

Expressivité limitée

Données « certaines »

Langage naturel

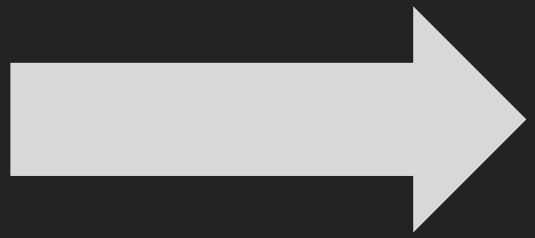
Expressivité
Communication
Fluidité

Données structurées

Réutilisation
ultérieure

Langage naturel

Expressivité
Communication
Fluidité



Données structurées

Réutilisation
ultérieure

COVID = **maladie émergente**

absence de données structurées **spécifiques**

Besoin de réponse **rapide**

➔**Traitement automatique** des documents cliniques pour

extraire données **ciblées**

extraire données **sans a priori**

Application à l'effet des traitements par **inhibiteurs calciques** chez les patients COVID

Utilité du TAL **souvent suggérée** [1,2]
mais **jamais en temps réel**

Association **Inhibiteurs calciques et COVID** évoquée [3]
Jamais dans une grande **étude multicentrique**

[1]Chapman et al., Proceedings of 36th symposium on the interface : Computing science and statistics, 2012
[2]Elkin et al., Annals of Internal Medicine, 2012
[3]Zhang et al., medRxiv, 2020

Base EDS-COVID

39 hôpitaux de l'APHP

84966 patients au 4 mai 2020

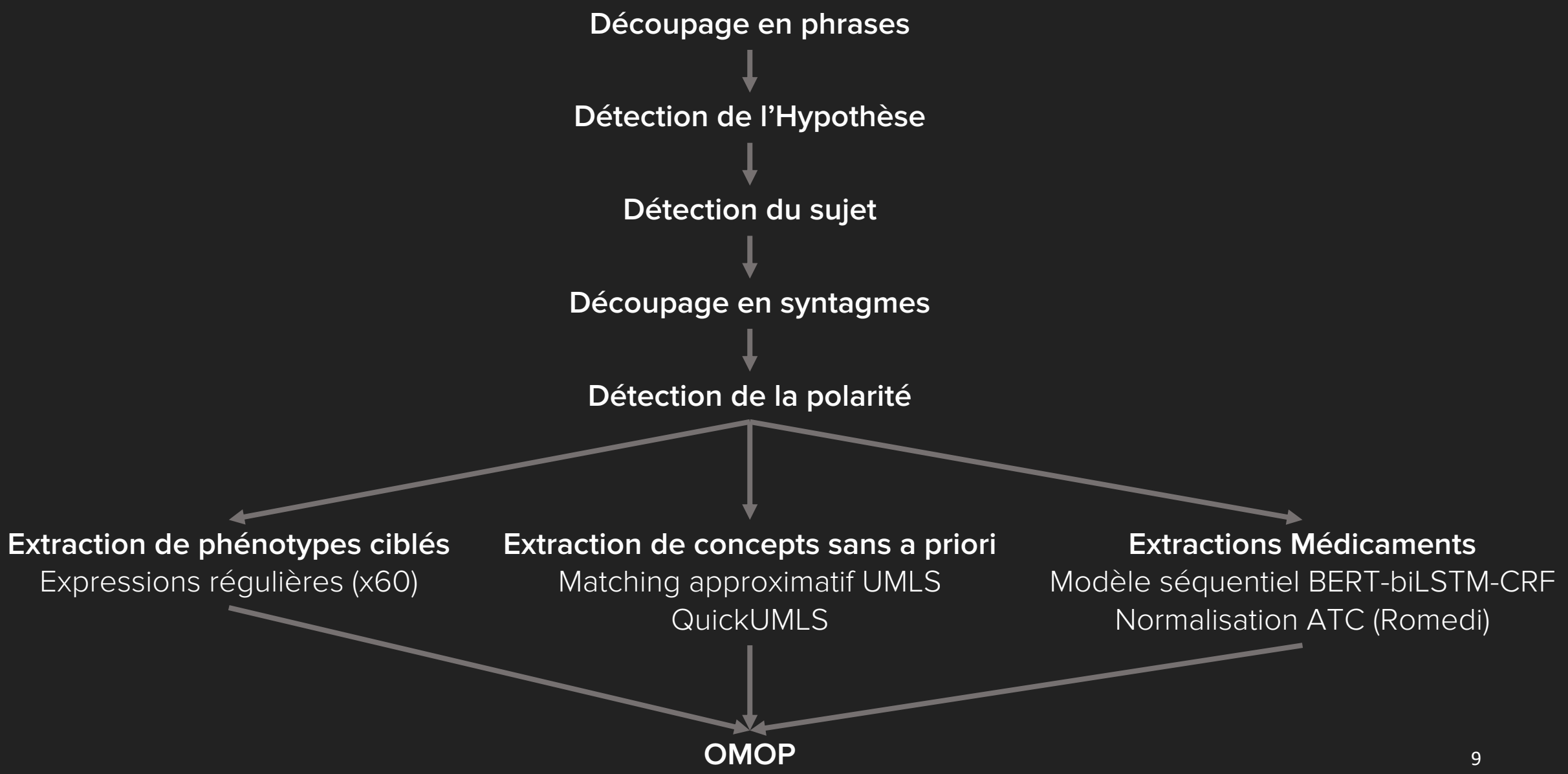
3 sources de données

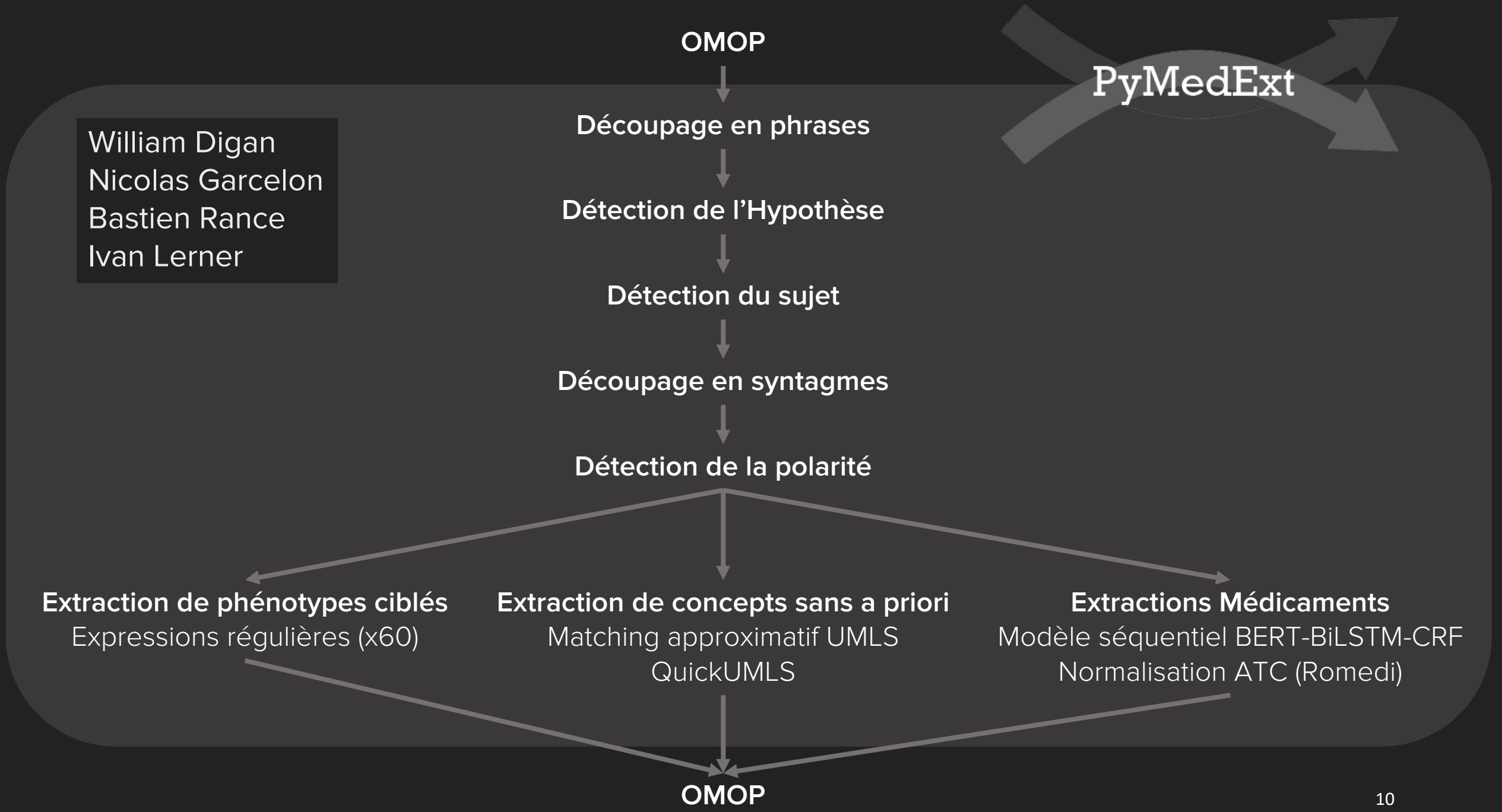
Codes diagnostics **CIM10**

Prescriptions structurées

Textes cliniques

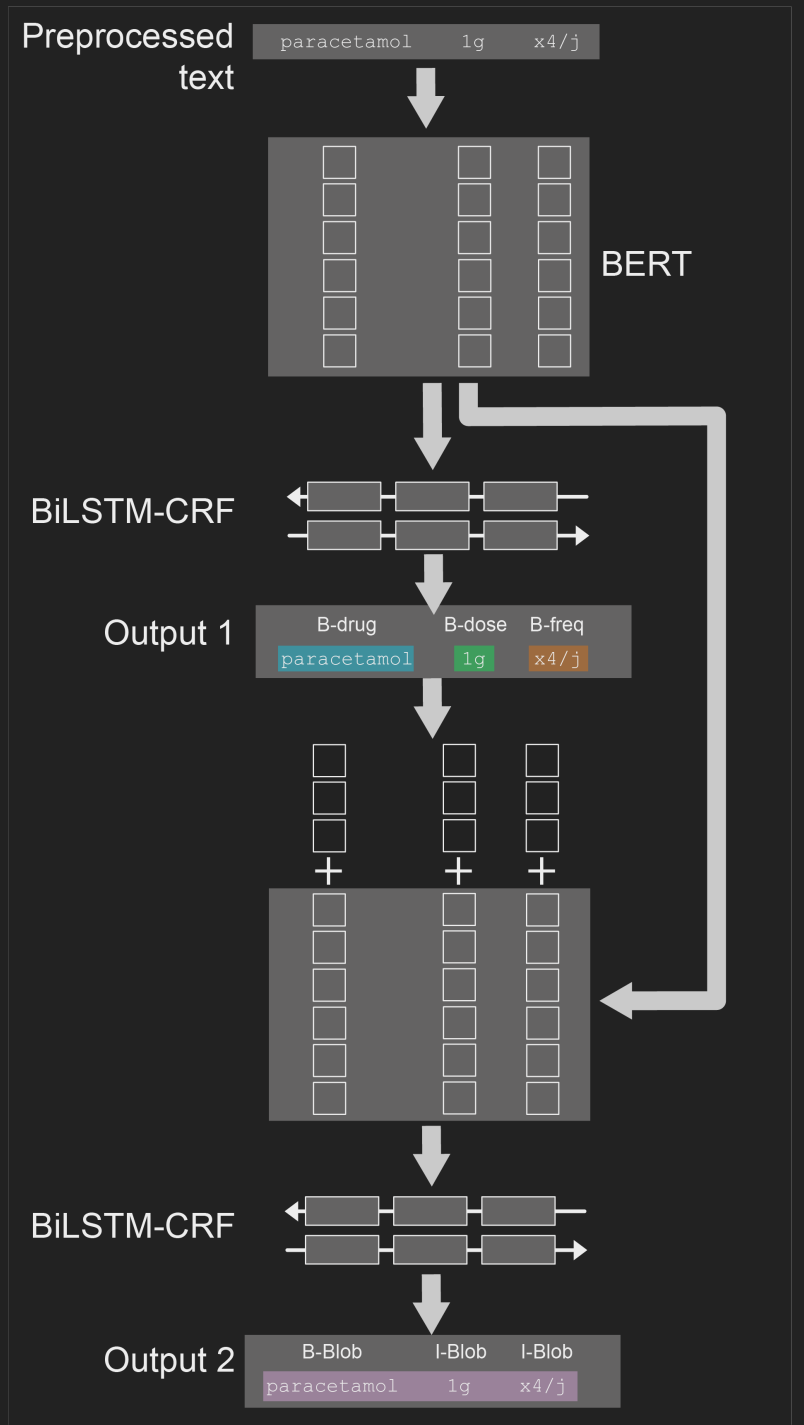
Description du pipeline d'extraction d'information déployé




OMOP
PyMedExt
William Digan
Nicolas Garcelon
Bastien Rance
Ivan Lerner
Découpage en phrases
Détection de l'Hypothèse
Détection du sujet
Découpage en syntagmes
Détection de la polarité
Extraction de phénotypes ciblés
Expressions régulières (x60)
Extraction de concepts sans a priori
Matching approximatif UMLS
QuickUMLS
Extractions Médicaments
Modèle séquentiel BERT-BiLSTM-CRF
Normalisation ATC (Romedi)
OMOP

Système séquentiel

BERT + BiLSTM-CRF

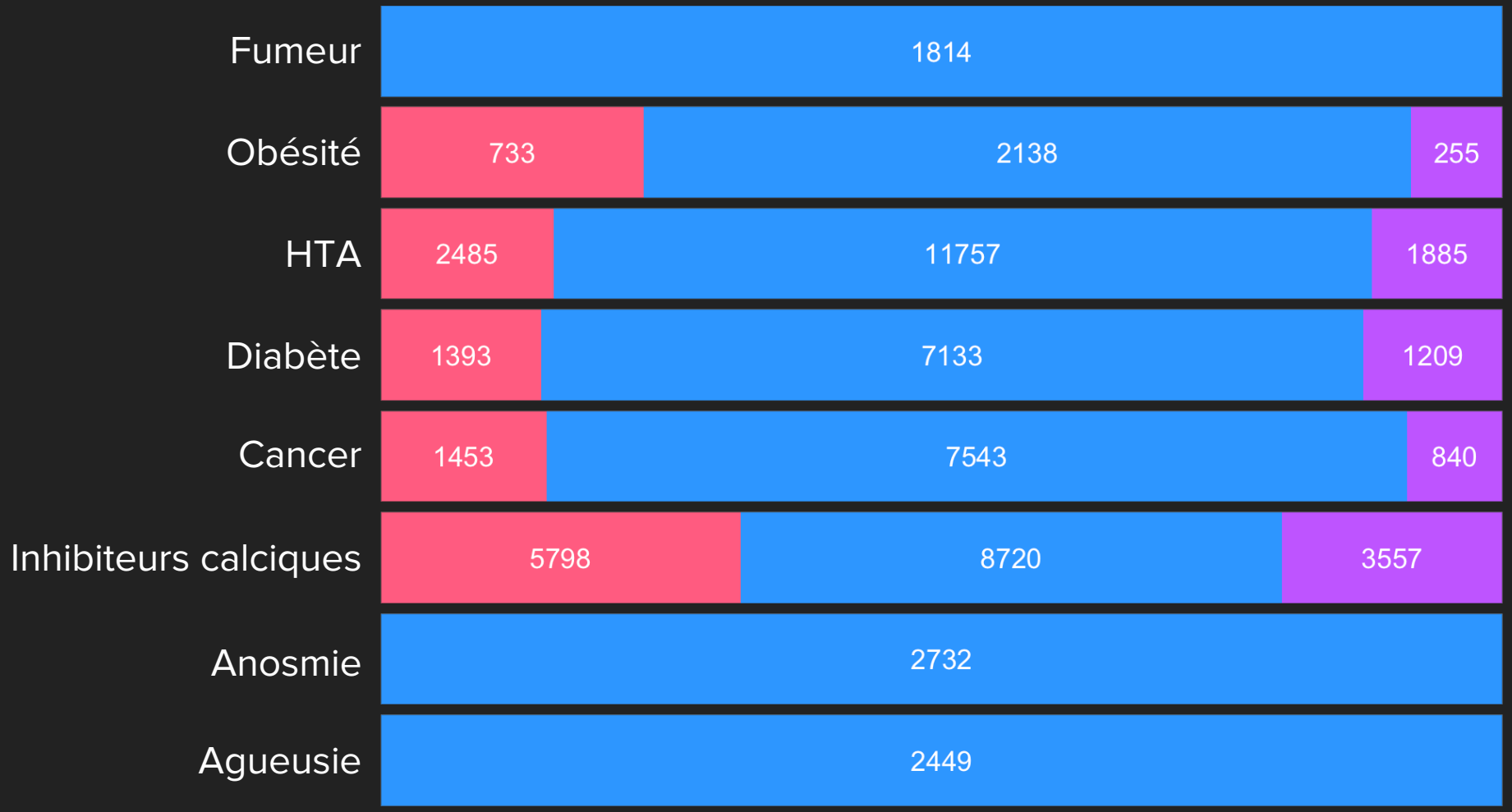


Ivan Lerner

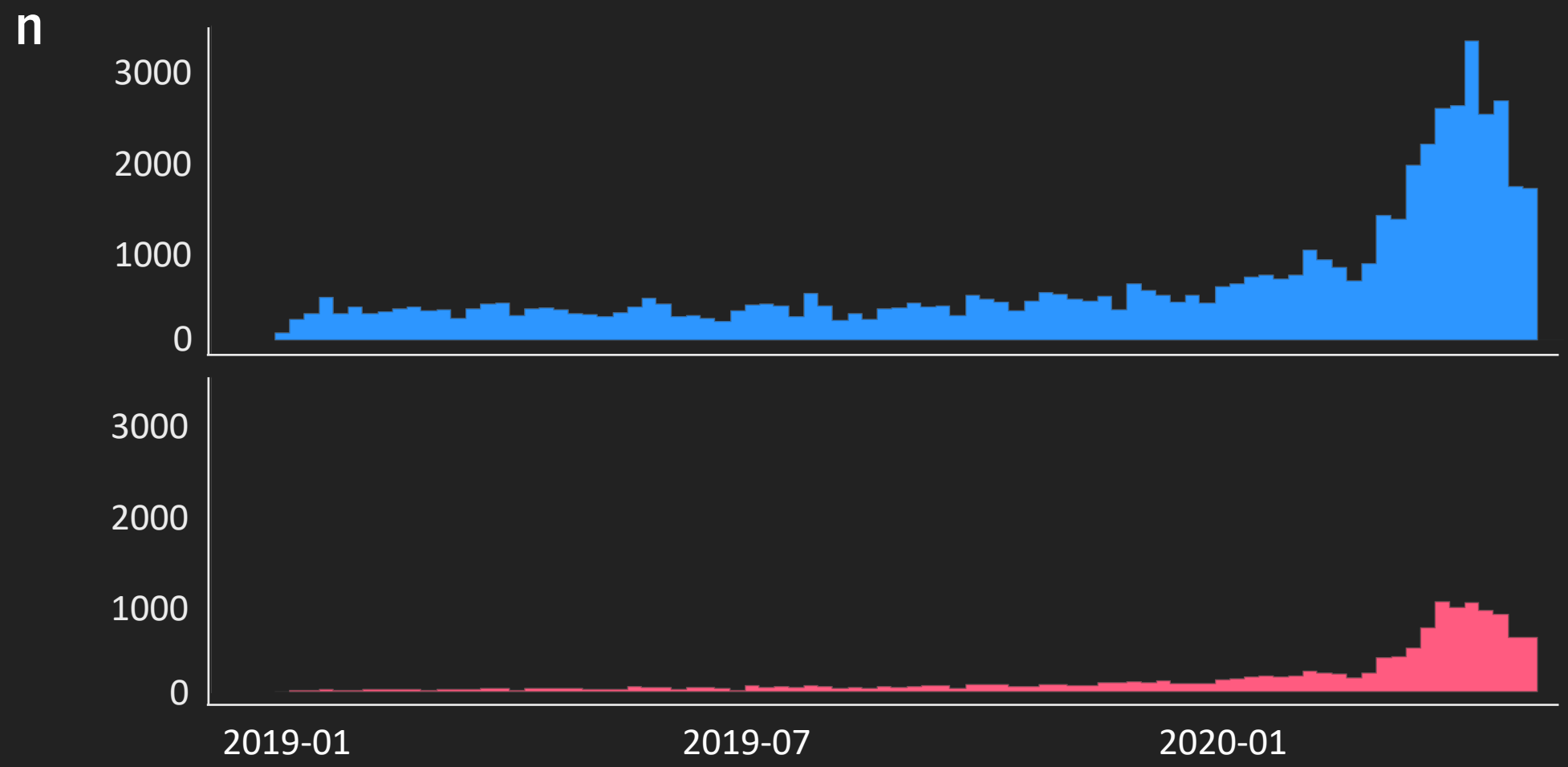
Le volume de données extraites du texte est supérieur aux données structurées

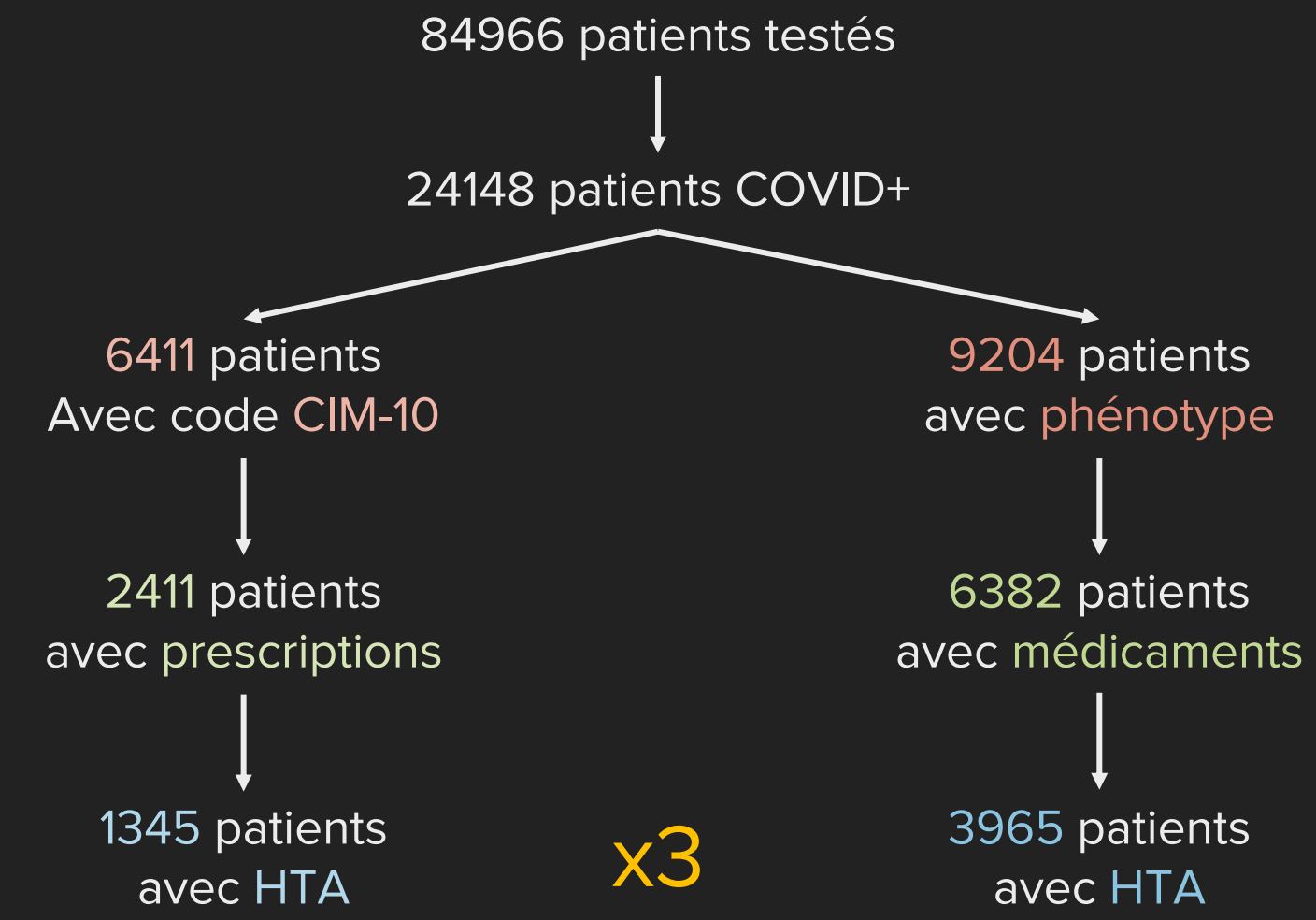
Structuré		Données extraites
médicaments	x7.2	médicaments médicaments médicaments médicaments médicaments médicaments médicaments médicaments médicaments
phénotypes	x15.2	phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes phénotypes

Nombre de patients présents dans les données structurées, les textes cliniques, les deux



Inhibiteurs calciques
dans les **textes cliniques** et dans les **données structurées**





Données structurées

TAL

	TAL N = 3965	Structuré N = 1343
Age		
18-44	175 (4.4%)	29 (2.2%)
45-64	1070 (27%)	205 (15%)
65-74	913 (23%)	252 (19%)
75-84	925 (23%)	392 (29%)
85+	882 (22%)	465 (35%)
Décès	810 (20%)	340 (25%)
Genre		
Homme	2236 (56%)	666 (50%)
Cancer	886 (22%)	444 (33%)
Diabète	1676 (42%)	560 (42%)
Obesité	518 (13%)	286 (21%)
Inhibiteurs calciques	1846 (47%)	525 (39%)

Analyse de survie

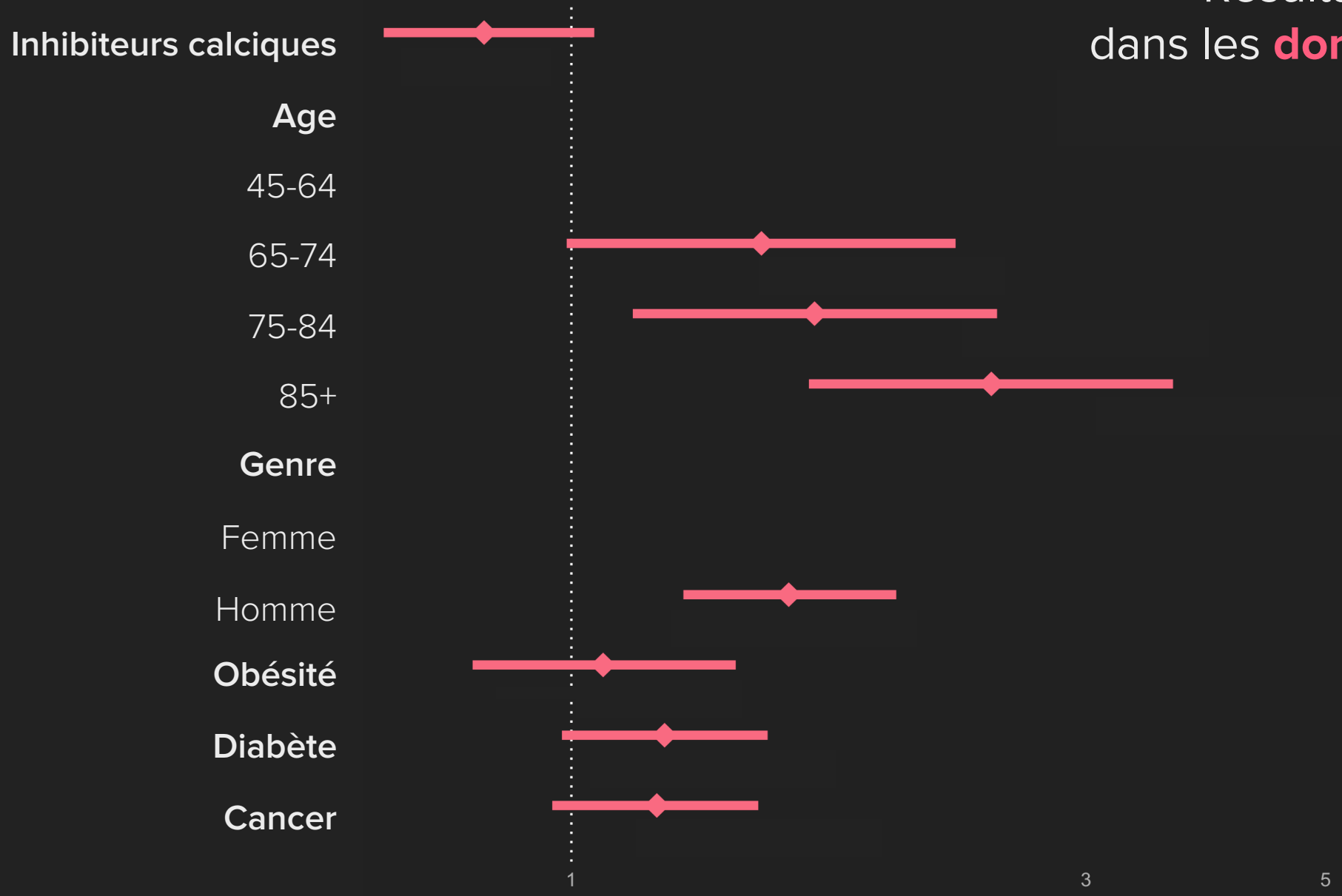
Risque de décès en fonction de variables explicatives
Données censurées

Modèle de COX multivarié

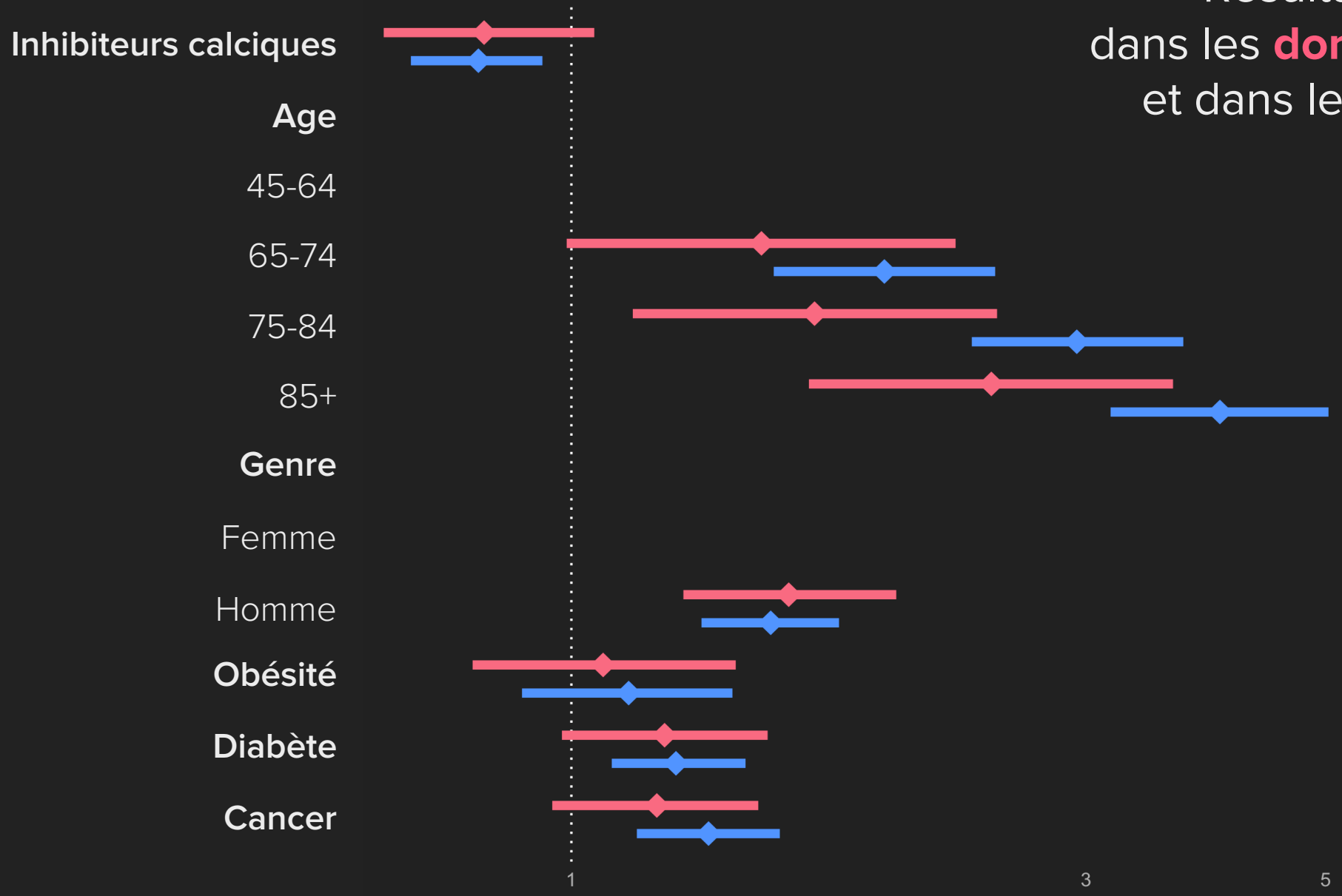
Plusieurs variables explicatives
Hypothèse des risques proportionnels

$$\lambda(t) = \lambda_0(t).\exp(Z_i^{\top}(t)\beta)$$

Résultats modèle de cox dans les **données structurées**



Résultats modèle de cox dans les **données structurées** et dans les **textes cliniques**



TAL utile pour une **maladie émergente**

- Informations **pertinentes**
- Déployé en **2 semaines**
- **Travaux pré-existants** (Médicaments, PyMedExt)
- **30+ projets de recherche** utilisent ces données
- Analyse MedWAS (Ivan Lerner)

Application clinique **à approfondir**

TAL utile pour une **maladie émergente**

- Informations **pertinentes**
- Déployé en **2 semaines**
- **Travaux pré-existants** (Médicaments, PyMedExt)
- **30+ projets de recherche** utilisent ces données
- Analyse MedWAS (Ivan Lerner)

Application clinique **à approfondir**

Neuraz, Antoine, Ivan Lerner, William Digan, Nicolas Paris, Rosy Tsopra, Alice Rogier, David Baudoin, et al. 2020. "Natural Language Processing for Rapid Response to Emergent Diseases : Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic."
Journal of Medical Internet Research 22 (8) : e20773. <https://doi.org/10.2196/20773>.

Merci pour votre attention !

David Baudoin
Anita Burgun
Leonardo Campillos
Kevin B Cohen
William Digan
Sarah F Feldman
Nicolas Garcelon
Jordan Jouffroy
Ivan Lerner
Nicolas Paris
Bastien Rance
Alice Rogier
Sophie Rosset
Rosy Tsopra