Exploitation des données de l'SNDS avec des techniques d'intelligence artificiel pour la recherche en cancérologie

M. Koume, M. Guyomard, AD. Bouhnik, L. Tassy, L. Seguin, **R. Urena**

Rencontres de SPF 2024, Paris (France)

SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université, Marseille, France

https://sesstim.univ-amu.fr/

Table of contents

Le SNDS

- Le Système National des Données de Santé (SNDS) est une **base de données exhaustive** regroupant des informations de santé sur l'ensemble de la population française.

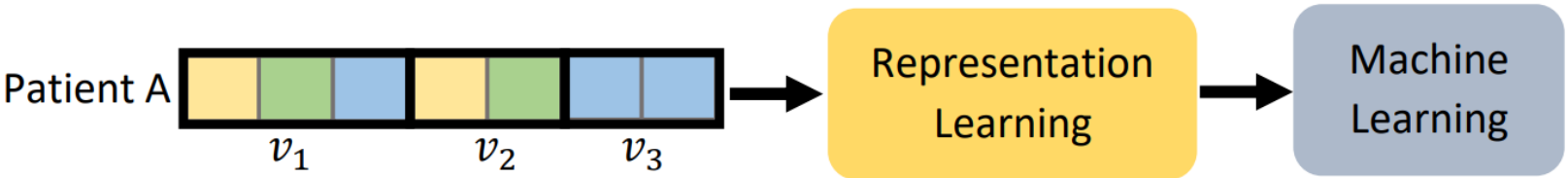
- Constitué de plusieurs sous-ensembles de données provenant de l'Assurance Maladie, des hôpitaux, et des registres de décès.

Composants du SNDS

- **Données de l'Assurance Maladie (SNIIRAM) :**
  - Informations sur les consultations médicales, hospitalisations, médicaments prescrits, et actes médicaux.
- **PMSI (Programme de Médicalisation des Systèmes d'Information) :**
  - Informations sur les séjours hospitaliers : diagnostics, actes pratiqués, et durée des séjours.
- **CepiDC (Centre d'épidémiologie sur les causes médicales de décès) :**
  - Données sur les causes de décès e
- **Données des registres nationaux :**
  - Registres des maladies chroniques, cancers, et autres pathologies spécifiques.

Challenges

- Dynamique temporelle
- Multi-modalité
- Données non structurées
- Très grande dimensionnalité



Études récents : SNDS en cancérologie

- **EDEN**, An Event DEtection Network for the annotation of Breast Cancer recurrences in administrative claims data[2]
- **FRESH**, The French Early Breast Cancer Cohort : A Resource for Breast Cancer Research and Evaluations of Oncology Practices Based on the French National Healthcare System Database (SNDS)[3].
- End-of life medical spending and care pathways in the last 12 months of life: A comprehensive analysis of the national claims database in France [6]

Prédiction de la peur de récidive du cancer du sein

- La peur de la récidive du cancer (PRC): "la peur ou l'inquiétude que le cancer revienne ou progresse dans le même organe ou dans une autre partie du corps"[7] .
- Préoccupation majeure : détérioration de la qualité de vie, fatigue, dépression, anxiété...
- Pas de mesure efficace en France pour identifier précocement les personnes susceptibles d'en souffrir.

Objectif

1. Proposer un modèle d'apprentissage interprétable pour identifier les patients à risque de développer une peur pathologique de la récidive.

2. Comprendre le lien entre la consommation de soins et la PRC.

3. Évaluer l'applicabilité à d'autres cancers (Prostate et colo-rectal).

Mesure de la Peur de la récidive

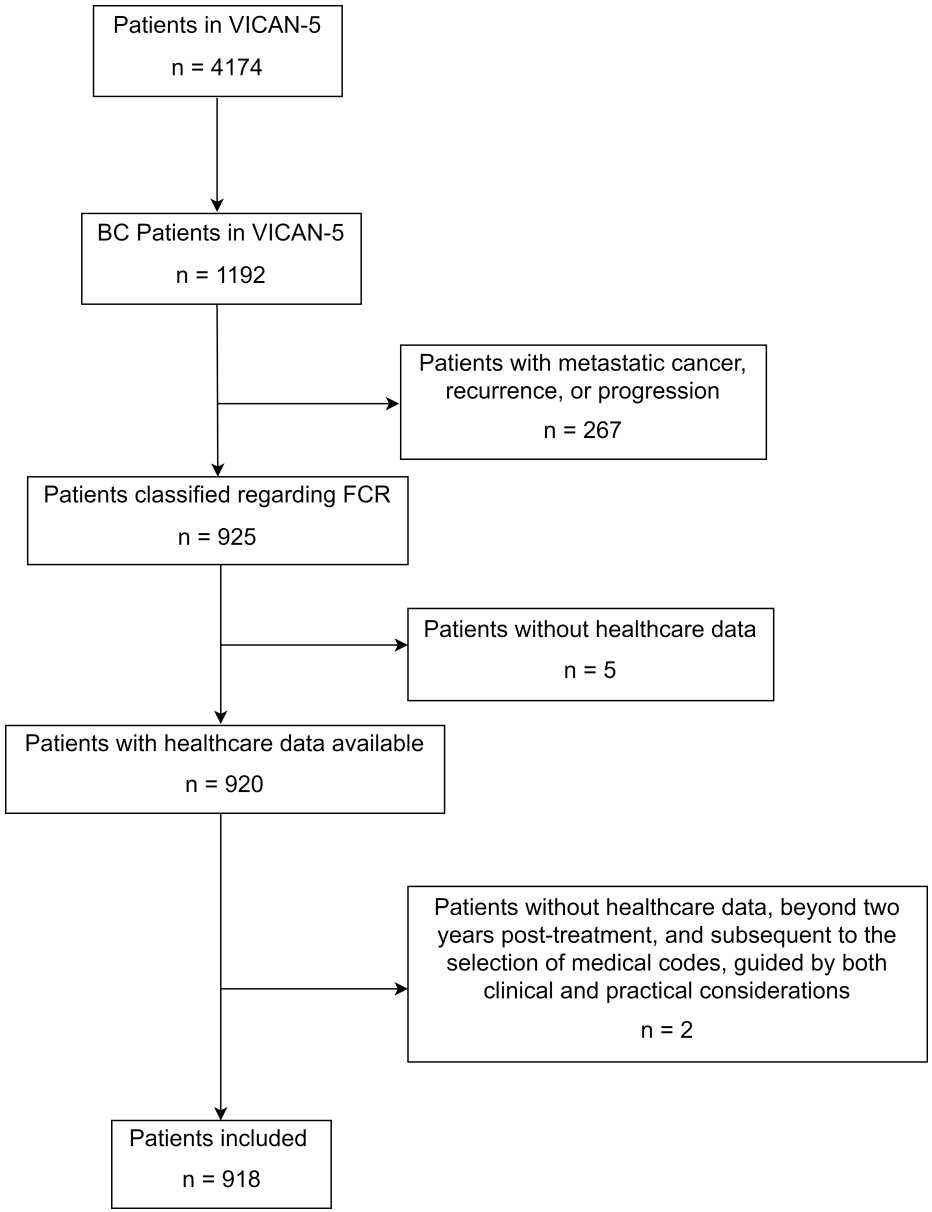
"À quelle fréquence pensez-vous à la possibilité d'une récidive de la maladie ?" sur une échelle de Likert à 5 points : jamais/quelques fois par mois/quelques fois par semaine/quelques fois par jour/plusieurs fois par jour.

Hypothèse

La PRC est liée à un surconsommation médicale, se traduisant par une utilisation fréquente de médicaments anxiolytiques et une sur-utilisation d'actes médicaux et biologiques.

Les données

- Enquêtes VICAN (VIe après le CANcer) : 2 ans et 5 ans après le diagnostic [1].
- Population : 20-85 ans au début de VICAN-2.
- Echantillon : 4174 individus, 12 localisations cancéreuses.
- Données : médicales (stade clinique, etc), patients (état de santé, séquelles, douleur, etc.) et sur la consommation de soins (extraction SNDS).


Patients in VICAN-5
n = 4174
BC Patients in VICAN-5
n = 1192
Patients with metastatic cancer, recurrence, or progression
n = 267
Patients classified regarding FCR
n = 925
Patients without healthcare data
n = 5
Patients with healthcare data available
n = 920
Patients without healthcare data, beyond two years post-treatment, and subsequent to the selection of medical codes, guided by both clinical and practical considerations
n = 2
Patients included
n = 918

Méthodologie : Pré-traitement des données

Processus de sélection guidé par des considérations cliniques et pratiques.

- Médicaments : effets analgésiques, psychotropes, contre l'anxiété ....
- Actes biologiques et médicaux pertinents : système sanguin, urinaire, circulatoire, respiratoire, musculosquelettique ...
- Prescriptions exclues : hormonothérapie, vaccins, autres facteurs externes incongrus avec l'investigation de la PRC.
- Nous avons exclu les deux premières années après le diagnostique.

Méthodologie : Pré-traitement des données

	MEDIC	BIO	CCAM	Total
Before selection				
# Patients (P)	920	905	916	920
# Medical codes	743	253	707	1703
# Medical prescriptions (E)	133 075	58 272	31 352	222 699
E/P	144.6	64.4	34.2	242.1
# Medical specialties	51	43	42	52
After selection				
# Patients (P)	913	900	914	918
# Medical codes	565	204	452	1221
# Medical prescriptions (E)	97 562	34 164	22 065	153 791
E/P	106	37.7	24.1	167.5
# Medical specialties	41	37	37	44

Méthodologie : Représentation des données patients

Soit un ensemble de n patients $P_{i=\{1,\ldots,n\}}$ où:

- $M = M_{i=\{1,\ldots,m\}}$, médicaments.
- $B = B_{i=1,\ldots,t}$, actes biologiques.
- $AM = AM_{i=\{1,\ldots,q\}}$ représente les actes médicaux.

Patient ID	M_1	M_2	M_3	...	B_1	B_2	B_3	...	AM_1	AM_2	AM_3	...
1	m_{11}	m_{12}	m_{13}	...	b_{11}	b_{12}	b_{13}	...	am_{11}	am_{12}	am_{13}	...
2	m_{21}	m_{22}	m_{23}	...	b_{21}	b_{22}	b_{23}	...	am_{21}	am_{22}	am_{23}	...
...												
n-1	$m_{(n-1)1}$	$m_{(n-1)2}$	$m_{(n-1)3}$	...	$b_{(n-1)1}$	$b_{(n-1)2}$	$b_{(n-1)3}$	...	$am_{(n-1)1}$	$am_{(n-1)2}$	$am_{(n-1)3}$	...
n	m_{n1}	m_{n2}	m_{n3}	...	b_{n1}	b_{n2}	b_{n3}	...	am_{n1}	am_{n2}	am_{n3}	...

Méthodologie :Pre processement des données

- **Sélection des variables:** Recursive Feature Elimination (RFE), Meta-Transformer- based feature selection (SFM), and Relief-based (ReliefF).

- **Rééquilibrage des données** (FCR 37%) : SMOTE et ADSYM.

Méthodologie : Apprentissage

- Données de entraînement : 80 %, Donnés de test : 20 %
- Gridsearch pour la selection des hyperparamètres.
- 3-fold cross validation

Résultats

Performance des modèles

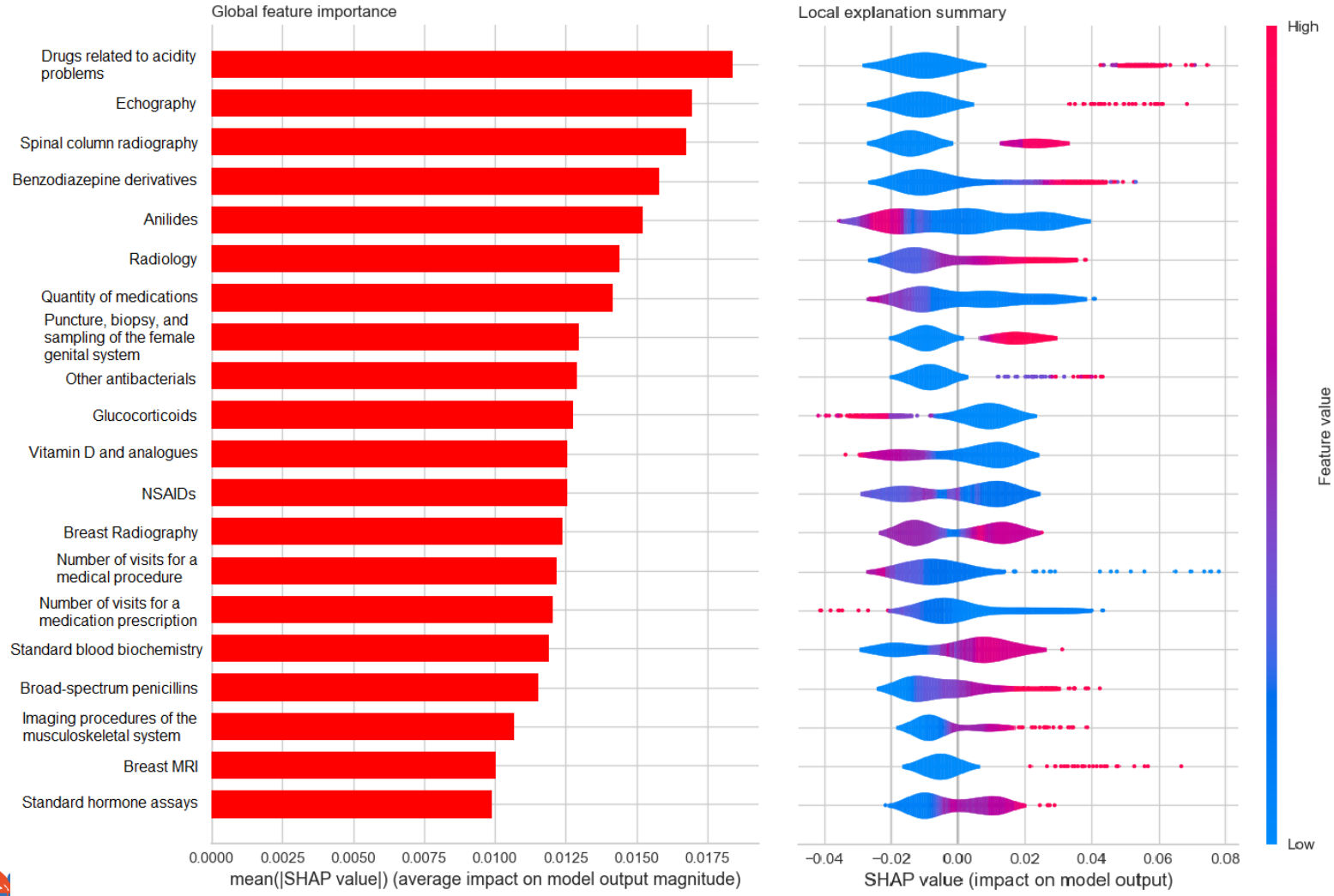
Models	AUC	Sensibilité	Précision	F1-score
GB	**0.66**	**0.70**	**0.71**	**0.71**
RF	0.64	0.60	0.72	0.65
XGB	0.62	0.65	0.67	0.66
MLP	0.66	0.63	0.75	0.68
LSTM	0.56	0.65	0.67	0.66
Bi-LSTM	0.57	0.78	0.65	0.71

Table: Comparaison de la prédiction des modèles

- Mamoudou Koume, Lorène Seguin, Anne-Déborah Bouhnik, and Raquel Urena. Predicting Fear of Breast Cancer Recurrence from Healthcare Reimbursement Data using Deep Learning.
  In *Proceedings of the IEEE Computer-Based Medical Systems (CBMS)*, Guadalajara, Mexico, 2024
- Mamoudou Koume, Lorène Seguin, Julien Mancini, Marc-Karim Bendiane, Anne-Déborah Bouhnik, and Raquel Urena.

  Predicting Fear of Breast Cancer Recurrence in women five years after diagnosis using Machine Learning and Healthcare Reimbursement Data from the French nationwide VICAN survey.
  Biomedical Engineering, 2024

Résultats


Global feature importance
Drugs related to acidity problems
Echography
Spinal column radiography
Benzodiazepine derivatives
Anilides
Radiology
Quantity of medications
Puncture, biopsy, and sampling of the female genital system
Other antibacterials
Glucocorticoids
Vitamin D and analogues
NSAIDs
Breast Radiography
Number of visits for a medical procedure
Number of visits for a medication prescription
Standard blood biochemistry
Broad-spectrum penicillins
Imaging procedures of the musculoskeletal system
Breast MRI
Standard hormone assays
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175
mean(|SHAP value|) (average impact on model output magnitude)
Local explanation summary
−0.04 −0.02 0.00 0.02 0.04 0.06 0.08
SHAP value (impact on model output)
High
Feature value
Low

Résultats

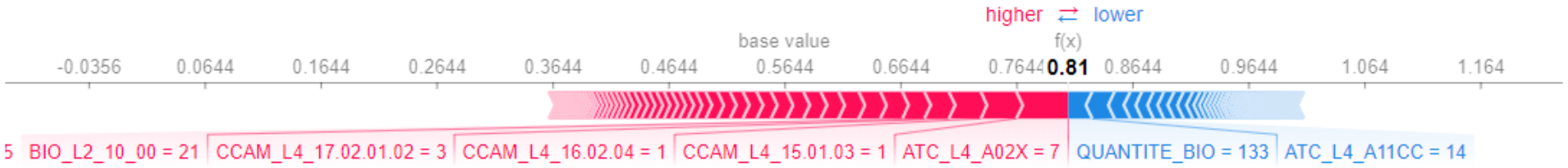


Figure: *

(A)

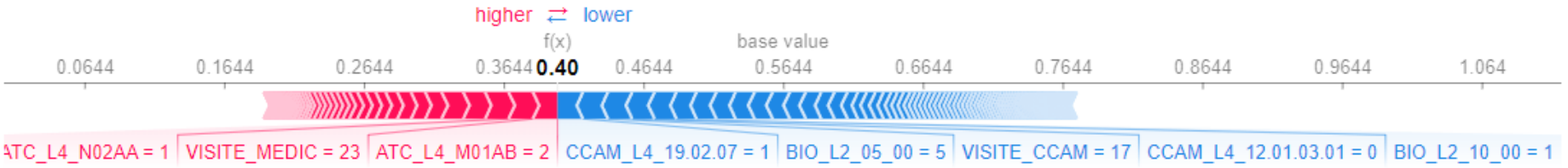


Figure: *

(B)

Figure: Interprétation des résultats au niveau individus. (A) Effet prédit pour une observation de la classe PRC. (B) Effet prédit pour une observation unique de la classe Non-PRC

Travail en cours

- Exploiter des données médico-administratives non étiquetées (6000 patients)
- Capturer la nature dynamique des trajectoires de soins et mieux comprendre les motifs évolutifs associés à la PRC.
- Utilisation du plan personnalisé de surveillance du cancer du sein saisir les changements dans les profils des patients au fil du temps.

Inclusion de la temporalité

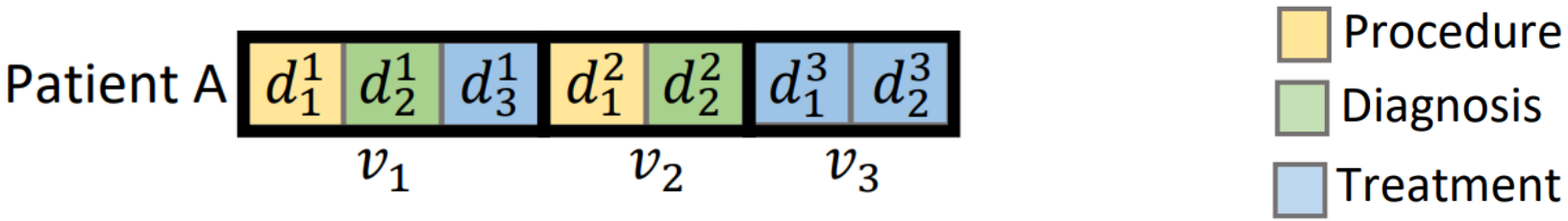


Figure: An example of EHR.

Procedure (CCAM)
Diagnosis (ICD-10)
Treatment (ATC)

- $V = \{v_1, \ldots, v_n\}$;
- j-th visit: $v_j = \{d_1^j, d_2^j, \ldots, d_{k_j}^j\}$;
- $v_j \subseteq \mathcal{C}, \quad \mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$;
- $L = \sum_{t=1}^{n} |v_t|$.

$n = 3$
$k_1 = 3, \; k_2 = k_3 = 2$

$L = 7$

Definition (Representation Learning Task)

Patient Representation Learning task involves extracting meaningful information from the dense mathematical representation of a patient within an embedding space or latent space

$$f_C : \mathbb{R}^L \rightarrow \mathbb{R}^m. \tag{1}$$

3 stratégies

- Natural Language Processing
- Autoencoders
- Transformers

Conclusion et discussion

1. **Originalité des données de VICAN**
   - Aucune étude antérieure n'a été menée, à notre connaissance, pour prédire la PRC en utilisant ce type de données de remboursement.
2. **Résultats prometteurs : preuve de concept**
   - Compétitifs, soulignant leur utilité potentielle dans l'identification des patients à risque de PRC clinique
3. **Implications**
   - Mise en œuvre de programmes de dépistage à l'échelle nationale, coordonnés par l'assurance maladie ou d'autres instituts de santé publique.

Réferences

AD Bouhnik, MK Bendiane, S Cortaredona, L Sagaon Teyssier, D Rey, C Bérenger, and V Seror.

The labour market, psychosocial outcomes and health conditions in cancer survivors: Protocol for a nationwide longitudinal survey 2 and 5 years after cancer diagnosis (the VICAN survey).
BMJ Open, 5:e005971, 2015.

Elise Dumas, Anne-Sophie Hamy, Sophie Houzard, Eva Hernandez, Aullène Toussaint, Julien Guerin, Laetitia Chanas, Victoire de Castelbajac, Mathilde Saint-Ghislain, Beatriz Grandal, Eric Daoud, Fabien Reyal, and Chloé-Agathe Azencott.

EDEN : An Event DEtection Network for the annotation of Breast Cancer recurrences in administrative claims data, November 2022.
arXiv:2211.08077 [cs].

Elise Dumas, Lucie Laot, Florence Coussy, Beatriz Grandal Rejo, Eric Daoud, Enora Laas, Amyn Kassara, Alena Majdling, Rayan Kabirian, Floriane Jochum, Paul Gougis, Sophie Michel, Sophie Houzard, Christine Le Bihan-Benjamin, Philippe-Jean Bousquet, Judicaël Hotton, Chloé-Agathe Azencott, Fabien Reyal, and Anne-Sophie Hamy.
The French Early Breast Cancer Cohort (FRESH): A Resource for Breast Cancer Research and Evaluations of Oncology Practices Based on the French National Healthcare System Database (SNDS).
Cancers, 14(11):2671, May 2022.

Mamoudou Koume, Lorène Seguin, Anne-Déborah Bouhnik, and Raquel Urena.

Predicting Fear of Breast Cancer Recurrence from Healthcare Reimbursement Data using Deep Learning.
In *Proceedings of the IEEE Computer-Based Medical Systems (CBMS)*, Guadalajara, Mexico, 2024.

Mamoudou Koume, Lorène Seguin, Julien Mancini, Marc-Karim Bendiane, Anne-Déborah Bouhnik, and Raquel Urena.

Predicting Fear of Breast Cancer Recurrence in women five years after diagnosis using Machine Learning and Healthcare Reimbursement Data from the French nationwide VICAN survey.
Biomedical Engineering, 2024.

Arnaud Nze Ossima, Daniel Szfetel, Bénédicte Denoyel, Omar Beloucif, Joelle Texereau, Louis Champion, Jean François