

Séminaire interne du SESSTIM

Prédiction de la Peur de la Récidive du Cancer en utilisant le Machine Learning et des données de remboursement de soins

Mamoudou KOUME

Sous la supervision de Anne-Déborah BOUHNİK et Raquel URENA

Marseille, 12 avril 2024

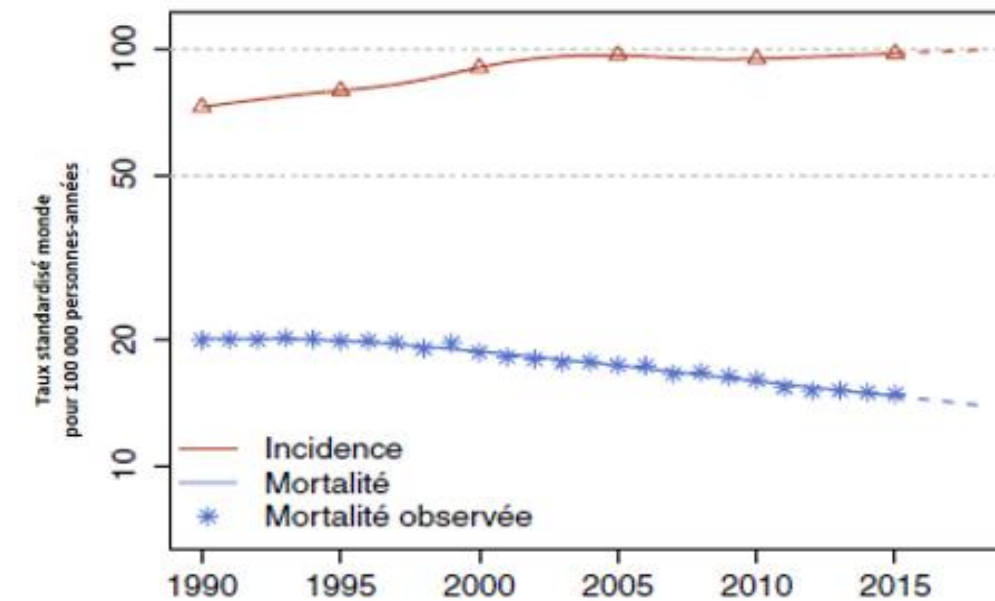


Sciences Economiques et Sociales de la
Santé & Traitement de l'Information Médicale
Inserm / IRD / Université AIX-MARSEILLE

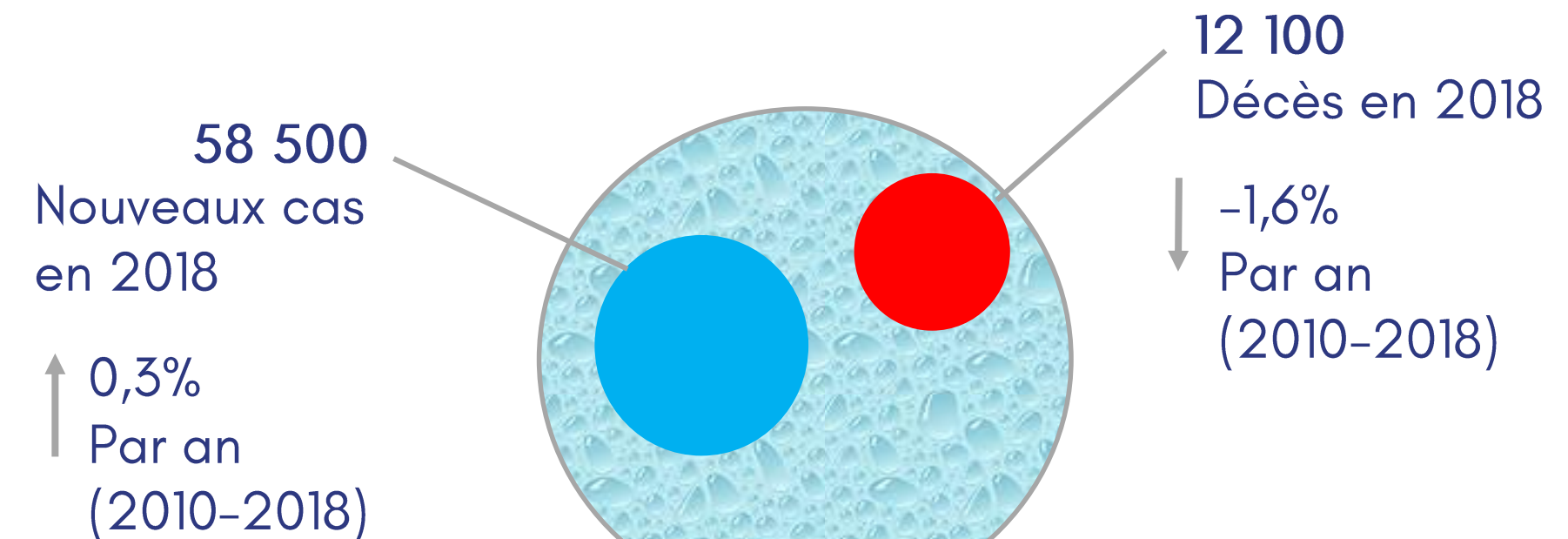
Inserm



- ▶ Cancer du sein : le plus fréquent et la principale cause de mortalité liée au cancer chez les femmes en France
- ▶ La Peur de la récurrence du cancer (PRC) définie comme "la peur ou l'inquiétude que le cancer revienne ou progresse dans le même organe ou dans une autre partie du corps".
- ▶ Préoccupation majeure : incertitude obsessionnelle sur l'avenir, détérioration de la qualité de vie, fatigue, dépression, anxiété et hyper-vigilance des symptômes corporels
- ▶ Des interventions psychosociales ciblées ont démontré leur efficacité dans la réduction de la PRC.
- ▶ Aucune mesure efficace n'existe en France pour identifier précocement les personnes susceptibles de souffrir de PRC



Source : Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018 - Volume 1 - Tumeurs solides.




Age médian au diagnostic : 63 ans

Taux de survie nette standardisée à 5 ans : 87%

Contexte (2/3)

Des études sur le sujet...

▶ Modèles de progression et de survie du cancer du sein avec des approches combinées de ML semi-explicables pour fournir une explication de l'importance des données démographiques, cliniques et biologiques,



COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL
journal homepage: www.elsevier.com/locate/csbj

Mini Review
Machine learning applications in cancer prognosis and prediction
Konstantina Kourou^a, Themis P. Exarchos^{a,b}, Konstantinos P. Exarchos^a, Michalis V. Karamouzis^c, Dimitrios I. Fotiadis^{a,b,*}

^a Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
^b IMBB – FORTH, Dept. of Biomedical Research, Ioannina, Greece
^c Molecular Oncology Unit, Department of Biological Chemistry, Medical School, University of Athens, Athens, Greece

▶ Identification de patients confrontés à un haut risque de mortalité à 180 jours avec des algorithmes de ML basés sur des dossiers électroniques de santé.

▶ Exploration de facteurs influençant le décès précoce chez les patients atteints de carcinome colorectal métastatique avec des modèles de ML et des données de remboursement.

▶ Prédiction du risque de cardiotoxicité dans une cohorte représentative au niveau national de patients atteints de cancer colorectal.

▶ Prédiction de saignements gastro-intestinaux liés aux antithrombotiques chez les patients ayant des antécédents médicaux spécifiques, en utilisant des données de remboursement médical et pharmaceutique des patients.

Support Care Cancer (2013) 21:941–949
DOI 10.1007/s00520-012-1609-2

ORIGINAL ARTICLE

Increased primary healthcare utilisation among women with a history of breast cancer

Carriene Roorda · Annette J. Berendsen · Feikje Groenhof · Klaas van der Meer · Geertruida H. de Bock

Journal of Cancer Survivorship
https://doi.org/10.1007/s11764-018-0714-8

Assessing the relationship between fear of cancer recurrence and health care utilization in early-stage breast cancer survivors

Amy K. Otto¹ · Emily C. Soriano¹ · Scott D. Siegel² · Stefanie T. LoSavio³ · Jean-Philippe Laurenceau¹

Received: 8 January 2018 / Accepted: 3 September 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

▶ Approche par ML bien que peu utilisée et appliquées à des outcomes non cliniques tels que la qualité de vie ou d'autres mesures de santé subjective, offrent également un potentiel inexploité pour l'identification de la PRC chez les survivants du cancer, associée à une surconsommation médicale.

Research

JAMA Oncology | Brief Report

Long-term Effect of Machine Learning-Triggered Behavioral Nudges on Serious Illness Conversations and End-of-Life Outcomes Among Patients With Cancer: A Randomized Clinical Trial

Christopher R. Manz, MD; Yichen Zhang, PhD; Kan Chen, MA; Qi Long, PhD; Dylan S. Small, PhD; Chalanda N. Evans, BS; Corey Chivers, PhD; Susan H. Regli, PhD; C. William Hanson, MD; Justin E. Bekelman, MD; Jennifer Braun, MHA, RN, BSN; Charles A. L. Rareshide, MS; Nina O'Connor, MD; Pallavi Kumar, MD, MPH; Lynn M. Schuchter, MD; Lawrence N. Shulman, MD; Mitesh S. Patel, MD, MBA; Ravi B. Parikh, MD, MPP

Cardiovascular Toxicology (2022) 22:130–140
https://doi.org/10.1007/s12012-021-09708-4

Using Machine Learning Approaches to Predict Short-Term Risk of Cardiotoxicity Among Patients with Colorectal Cancer After Starting Fluoropyrimidine-Based Chemotherapy

Chao Li¹ · Li Chen² · Chiahung Chou^{1,3} · Surachat Ngorsurachues¹ · Jingjing Qian¹

Construction and validation of nomograms combined with novel machine learning algorithms to predict early death of patients with metastatic colorectal cancer

Yalong Zhang^{1†}, Zunni Zhang^{2†}, Liuxiang Wei¹ and Shujing Wei^{1,*}

¹Department of Ultrasound Medicine, The Fifth Affiliated Hospital of Guangxi Medical University, Nanning, China. ²Department of Clinical Laboratory, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, China

JAMA Network | Open

Original Investigation | Gastroenterology and Hepatology

Comparative Effectiveness of Machine Learning Approaches for Predicting Gastrointestinal Bleeds in Patients Receiving Antithrombotic Treatment

Jeph Herrin, PhD; Neena S. Abraham, MD; Xiaoxi Yao, PhD; Peter A. Noseworthy, MD; Jonathan Inselman, MS; Nilay D. Shah, PhD; Che Ngufor, PhD



Objet de l'étude

Proposer un modèle de Machine Learning (ML) interprétable pour identifier les patients à risque de développer une PRC clinique suite au diagnostic de cancer.



Pertinence de l'étude

Comprendre le lien entre la consommation de soins et la PRC
Évaluer l'applicabilité à d'autres sites de cancer.



Question de recherche

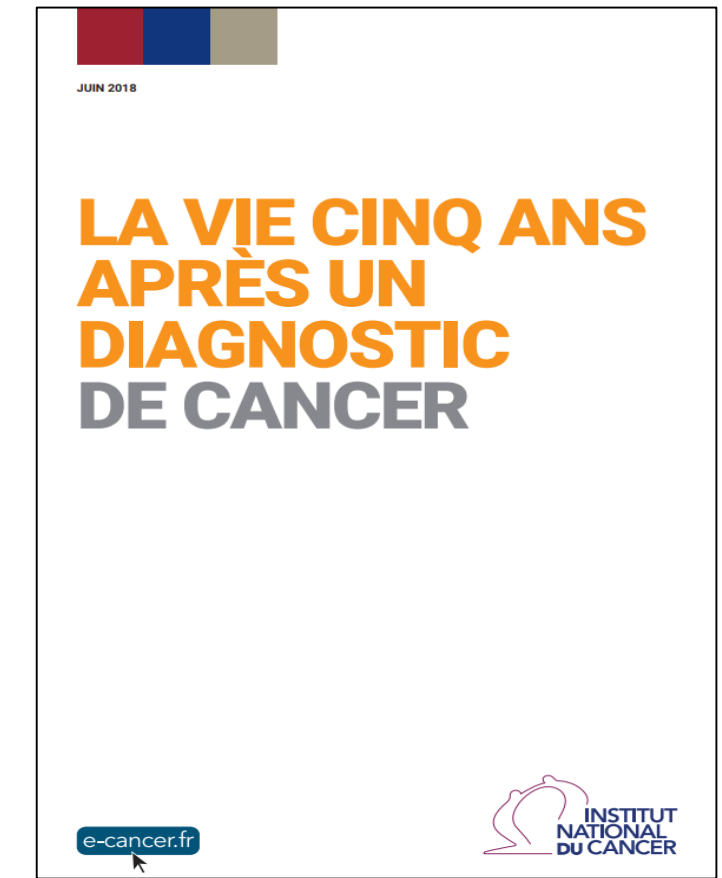
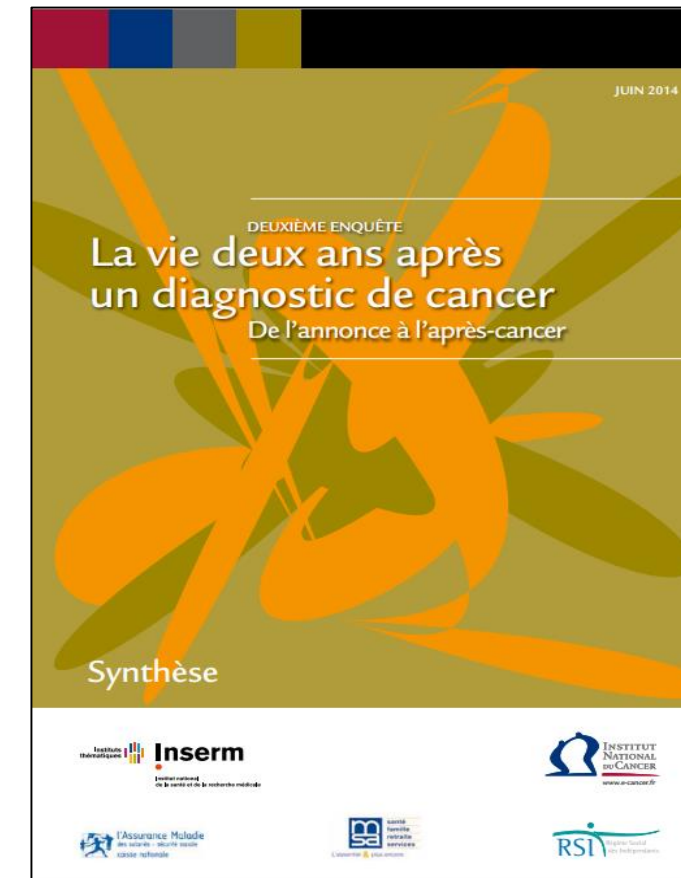
Comment l'application du ML aux données de consommation de soins peut-elle contribuer à identifier de façon précoce les patients présentant un risque élevé de PRC, et quelles avancées peut-elle apporter en termes de personnalisation des soins et d'amélioration des résultats de santé?



Notre hypothèse est que la PRC est associée à une surconsommation médicale, qui peut se manifester par une utilisation excessive de médicaments utilisés pour traiter l'anxiété, ainsi que par une surutilisation d'actes biologiques et médicaux.

Enquêtes VICAN (Vie après le CANcer) : Pluridisciplinarité, 3 sources, France métropolitaine

- ▶ VICAN-2 puis VICAN-5 : 2 ans et 5 ans après le diagnostic
- ▶ Population : 20-85 ans au début de VICAN-2
- ▶ Recueil des données : interviews téléphoniques, questionnaires auprès des soignants pour les données médicales, SNIIRAM (Système National d'Information Inter-Régimes de l'Assurance Maladie)
- ▶ Type de données :
 - ❑ Données médicales : stade clinique, grade, etc.
 - ❑ Données patients : état de santé, prise en charge, vie quotidienne, vie professionnelle, qualité de vie, fatigue, séquelles, douleur, sexualité, etc.
 - ❑ Données sur la consommation de soins : consultations médicales, délivrances de médicaments, hospitalisations, etc.
- ▶ VICAN-5 : 4174 individus, 12 localisations cancéreuses.



Dans VICAN-5, parmi les participants n'ayant signalé aucune rechute, la PRC a été mesurée à l'aide de la question suivante : "À quelle fréquence pensez-vous à la possibilité d'une récurrence de la maladie ?" avec des réponses 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. Les réponses ont été catégorisées en Non-PRC (jamais) et PRC (toutes les autres réponses) dans la présente étude.

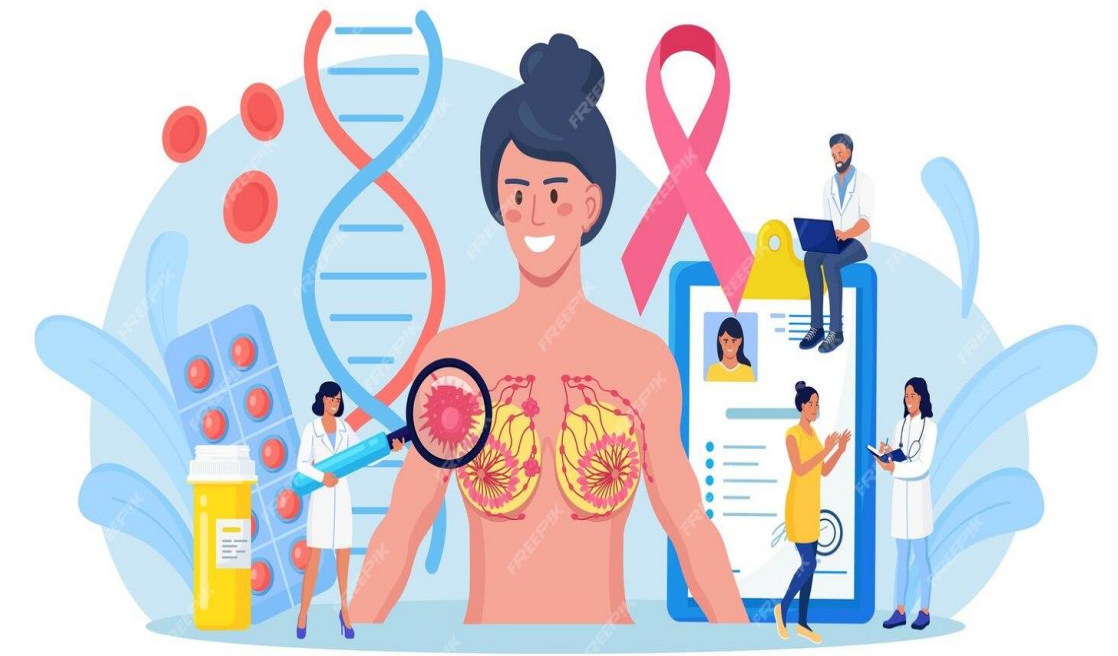
Données issues de VICAN

▶ Dans l'enquête VICAN-5, 6013 femmes diagnostiquées avec un cancer du sein non métastatique n'ont présenté aucun signe de récurrence ou de progression de la maladie cinq ans après le diagnostic.

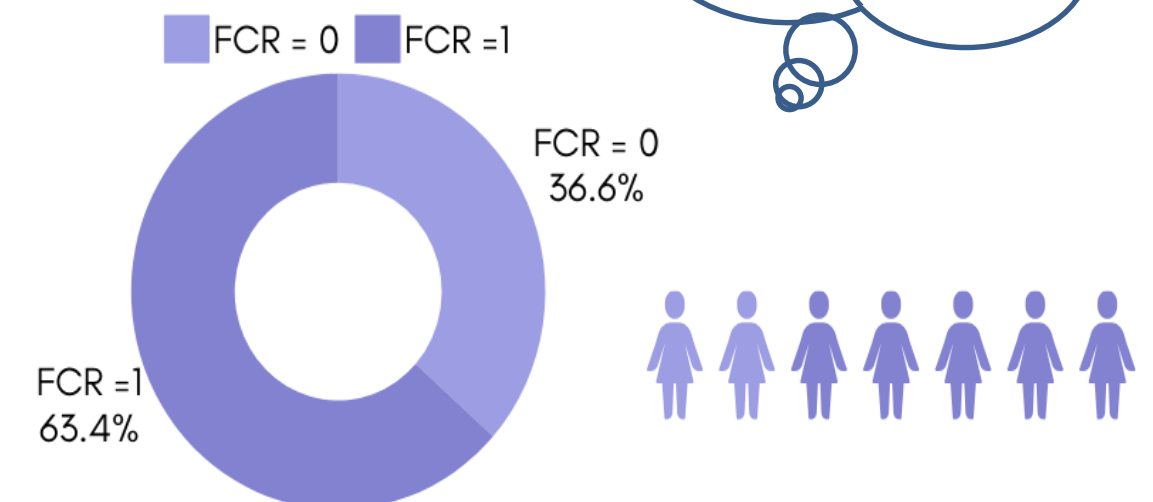
↳ Progression de la maladie : apparition de métastases, de nouveaux diagnostics de cancer, d'admissions en soins palliatifs, ou le recours à des traitements anticancéreux au moins 24 mois après le diagnostic.

▶ Sur cette cohorte, 930 ont activement participé à l'enquête VICAN-5 : entretiens téléphoniques, information sur la PRC, données de remboursement de soins de santé (médicaments, actes médicaux, tests de laboratoire, etc.).

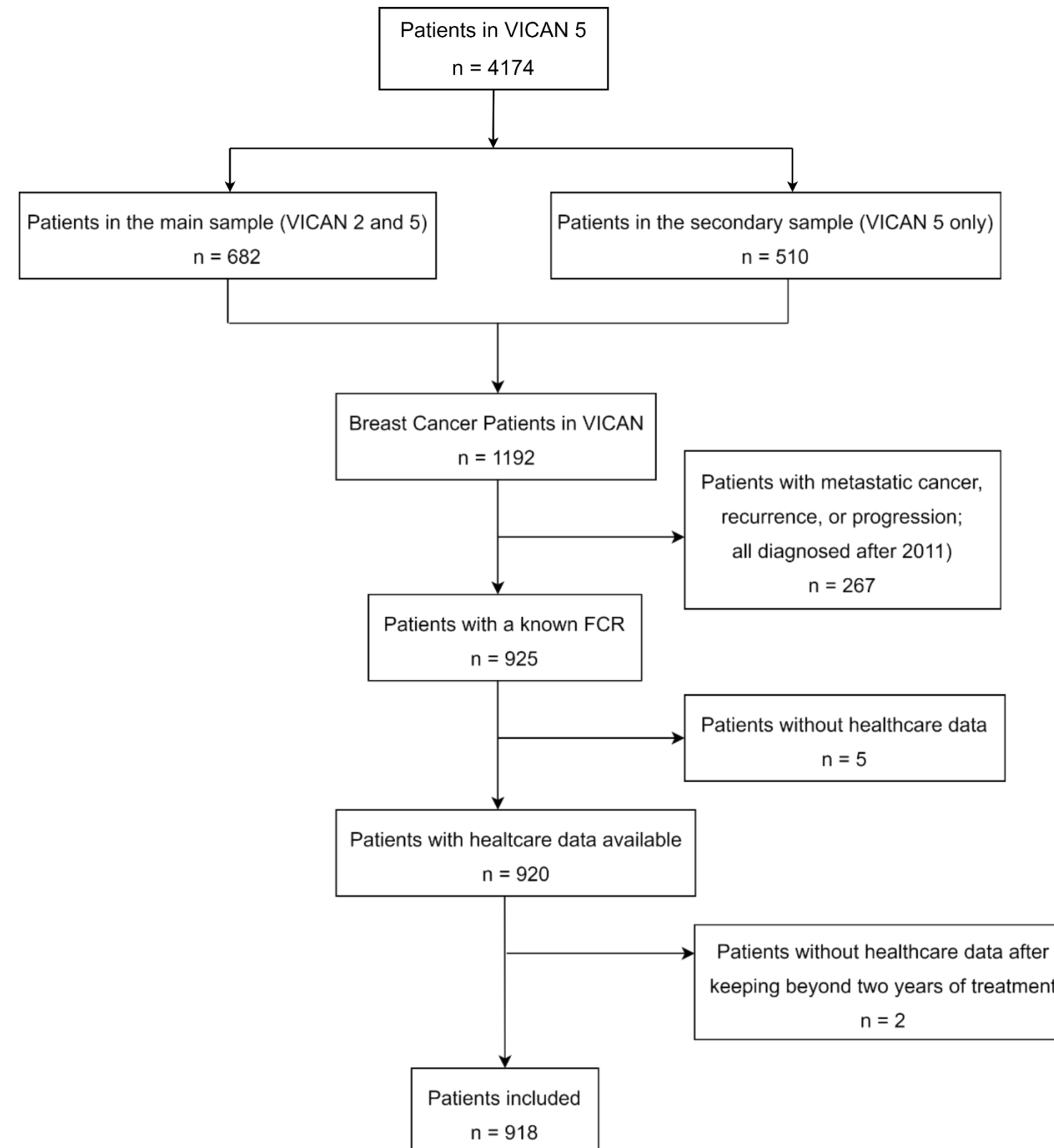
▶ Données de remboursement au-delà des deux premières années de traitement (phase aiguë du traitement).



Données assez déséquilibrées !



Patients inclus dans l'étude



01

Processus de selection 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.

02

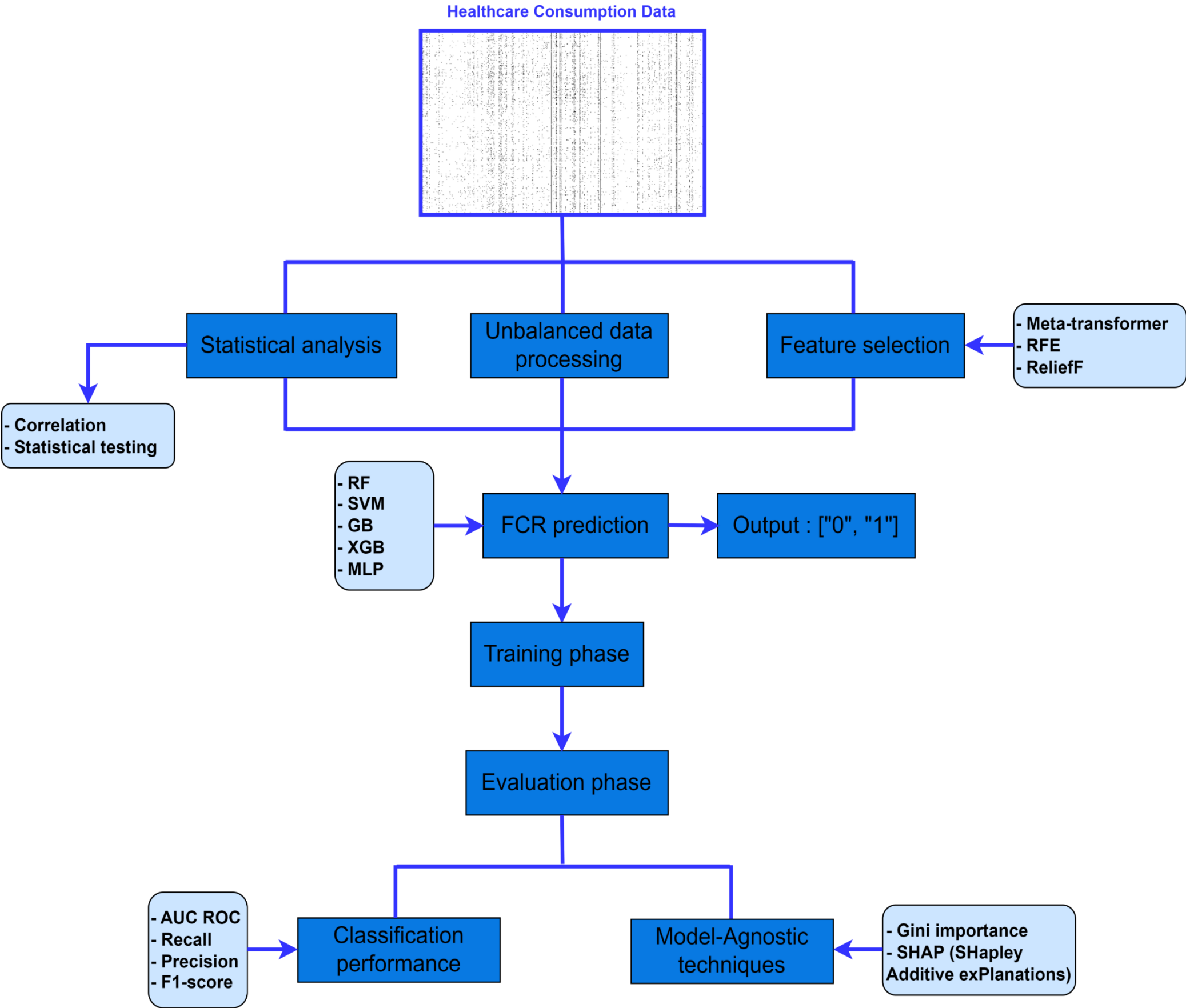
Analyse des données et identification des caractéristiques pertinentes.

- ▶ Analyses descriptives
- ▶ Tests statistiques, corrélation, clustering, etc
- ▶ Sélection de variables, suréchantillonnage, etc.

03

Modèles prédictifs.

- ▶ Méthodes ensemblistes: Random Forest, Gradient Boosting, XGBoost, etc.
- ▶ Autres modèles de classification : Support Vector Machine, Multilayer Perceptron.
- ▶ Évaluation, interprétation et validation des modèles proposés



Expérimentation 1	Expérimentation 2
Derniers niveaux de classification	Avant-derniers niveaux de classification
Niveau 5 pour les médicaments Niveau 5 pour les actes médicaux Niveau 3 pour les actes biologiques	Niveau 4 pour les médicaments Niveau 4 pour les actes médicaux Niveau 2 pour les actes biologiques

Résultats (1/5)

Analyse statistique

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

Table 1: Statistics on Healthcare Consumption tables before and after selection

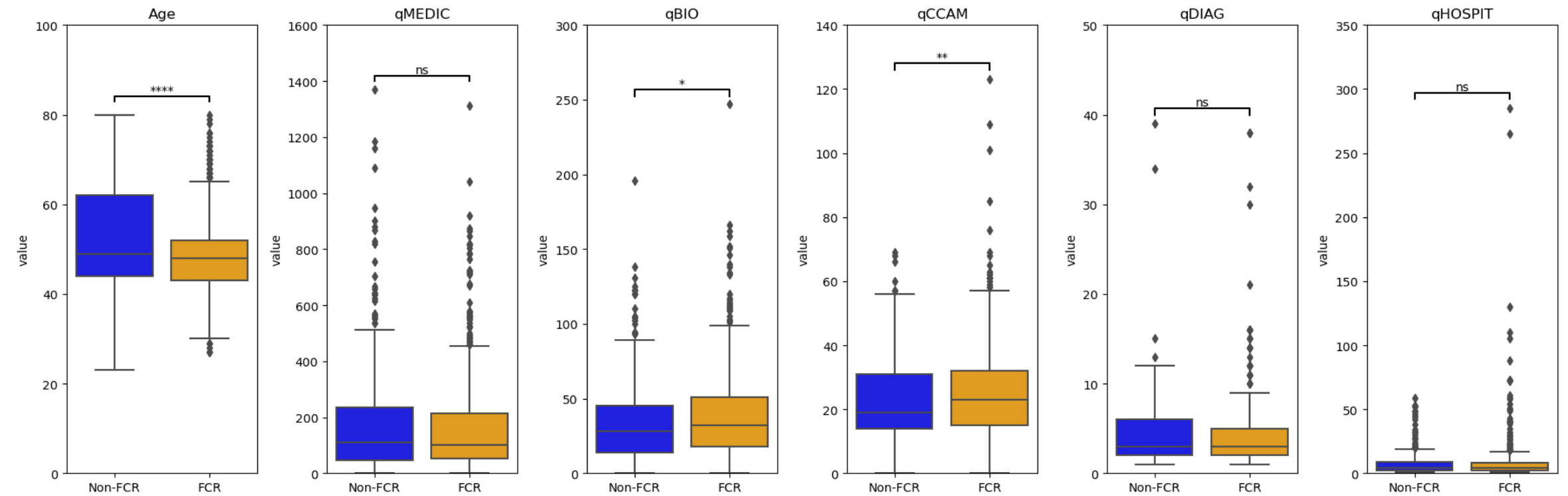


Figure. Characteristics of the study population. Boxplots illustrating the distribution of age and BC medical data, including medication prescriptions (qMEDIC), biological procedures (qBIO), medical procedures (qCCAM), hospitalization records (qHOSPIT), and diagnostic procedures (qDIAG) between FCR and Non-FCR groups. Statistical significance was determined using t-test, with 'ns' indicating non-significant differences and stars denoting the level of significance (* $p < \$ 0.05$, ** $p < \$ 0.01$, *** $p < \$ 0.001$, **** $p < \$ 0.0001$). The boxes represent the interquartile range (IQR), with median lines, whiskers indicating 1.5 times the IQR, and outliers displayed as individual points.

Relation entre un ensemble de n patients $P = \{1, \dots, n\}$ et un ensemble d'attributs $A = \{M, B, C\}$, où $M = M_i_{\{i=1, \dots, m\}}$ représente les médicaments, $B = B_i_{\{i=1, \dots, t\}}$ les actes biologiques et $C = C_i_{\{i=1, \dots, q\}}$ représente les actes médicaux.



Patient ID	M_1	M_2	M_3	...	B_1	B_2	B_3	...	C_1	C_2	C_3	...
1	X		X		X		X		X			
2			X		X				X		X	
3	X	X					X			X		
...												
n-1		X			X		X		X		X	
n		X		X	X	X	X	X		X		

Table 2: A Comprehensive Patient Profiles

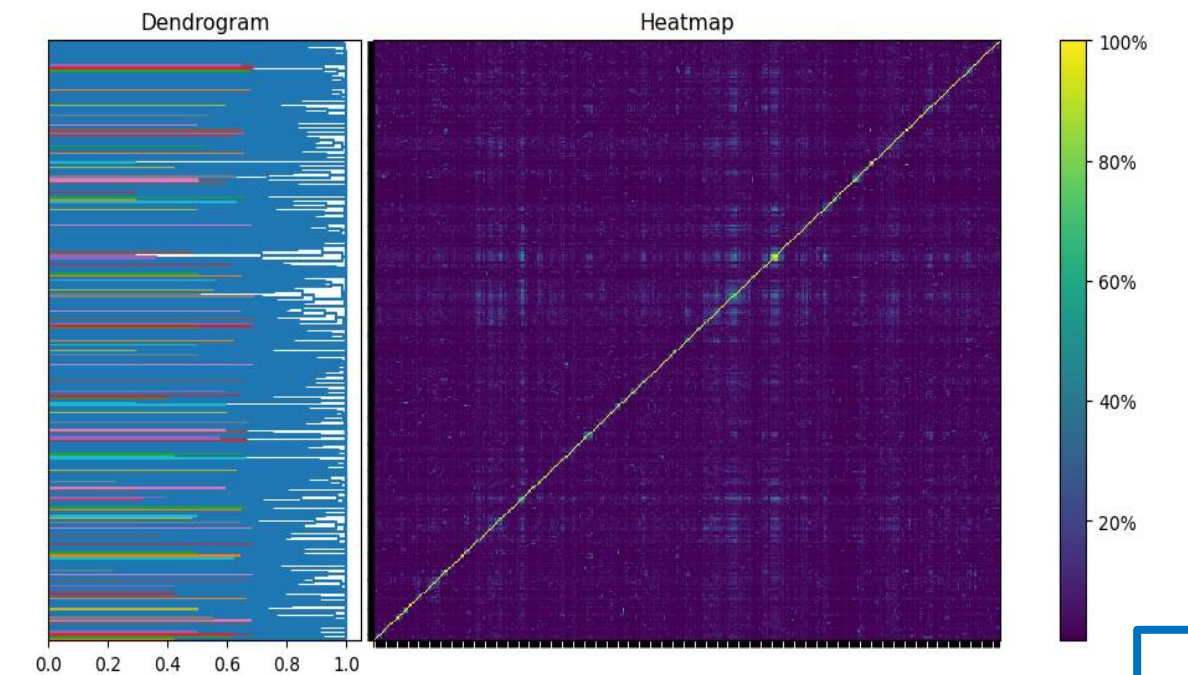
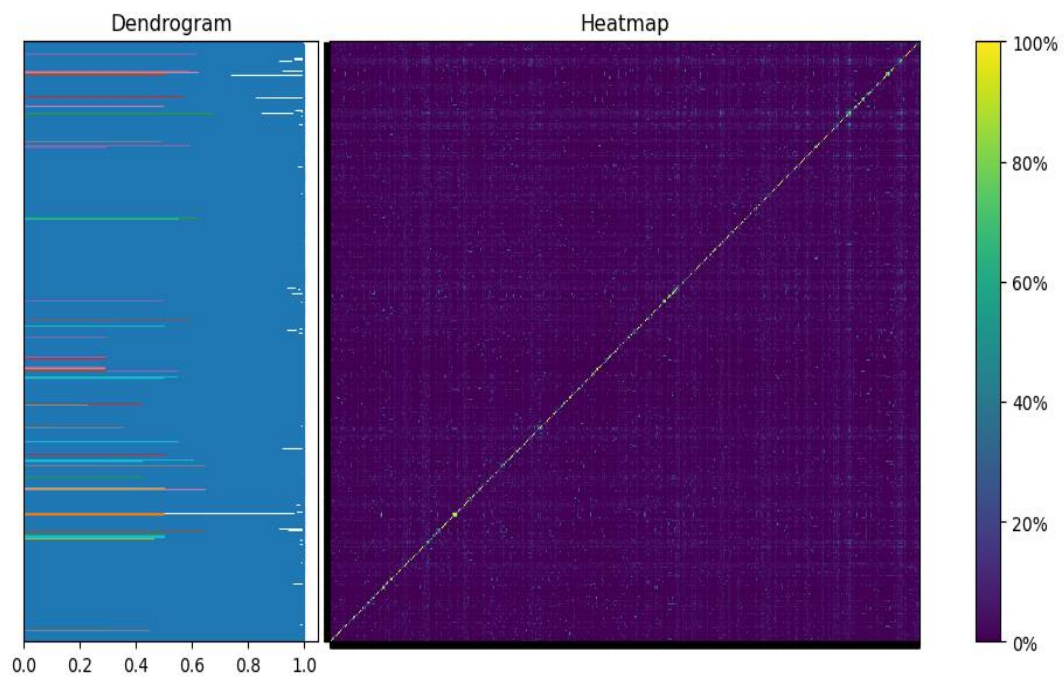
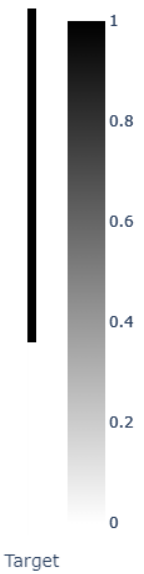
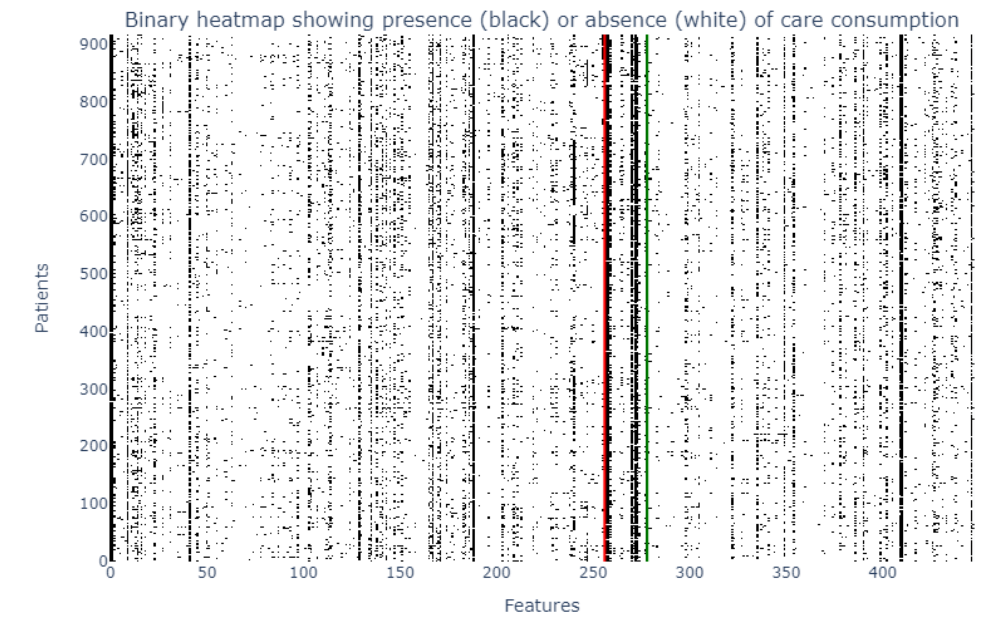
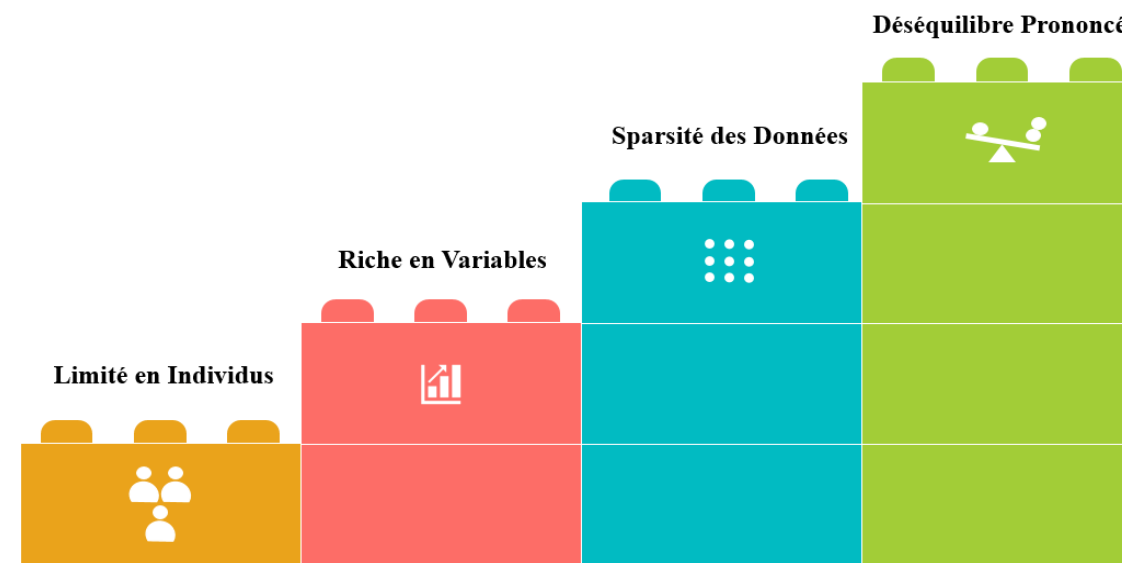
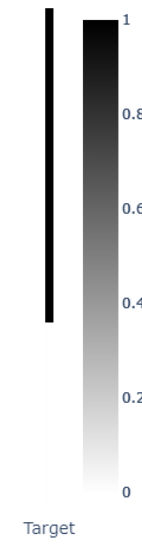
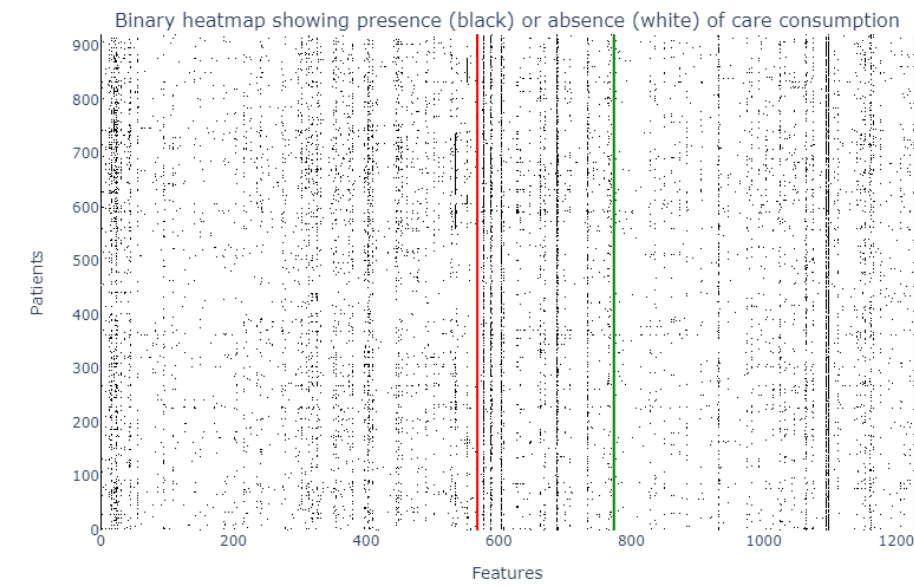
Résultats (2/5)

Analyse statistique

Expérimentation 1

Quand les données font du fun !

Expérimentation 2



Résultats (3/5)

Résultats des modèles

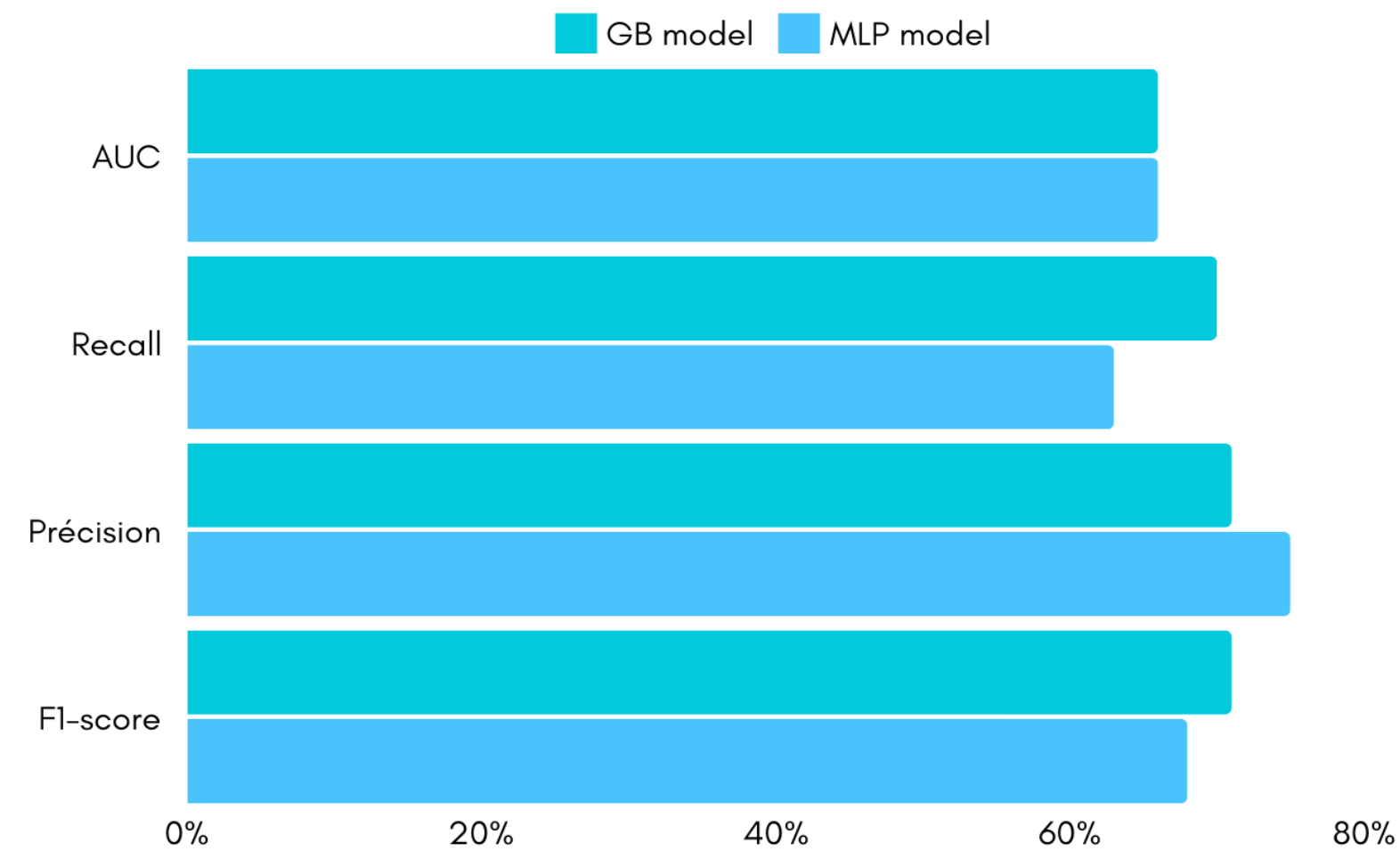


Figure. Meilleurs modèles obtenus dans l'expérience 2 : GB avec SFM et SMOTE et MLP avec ADASYN et ReliefF

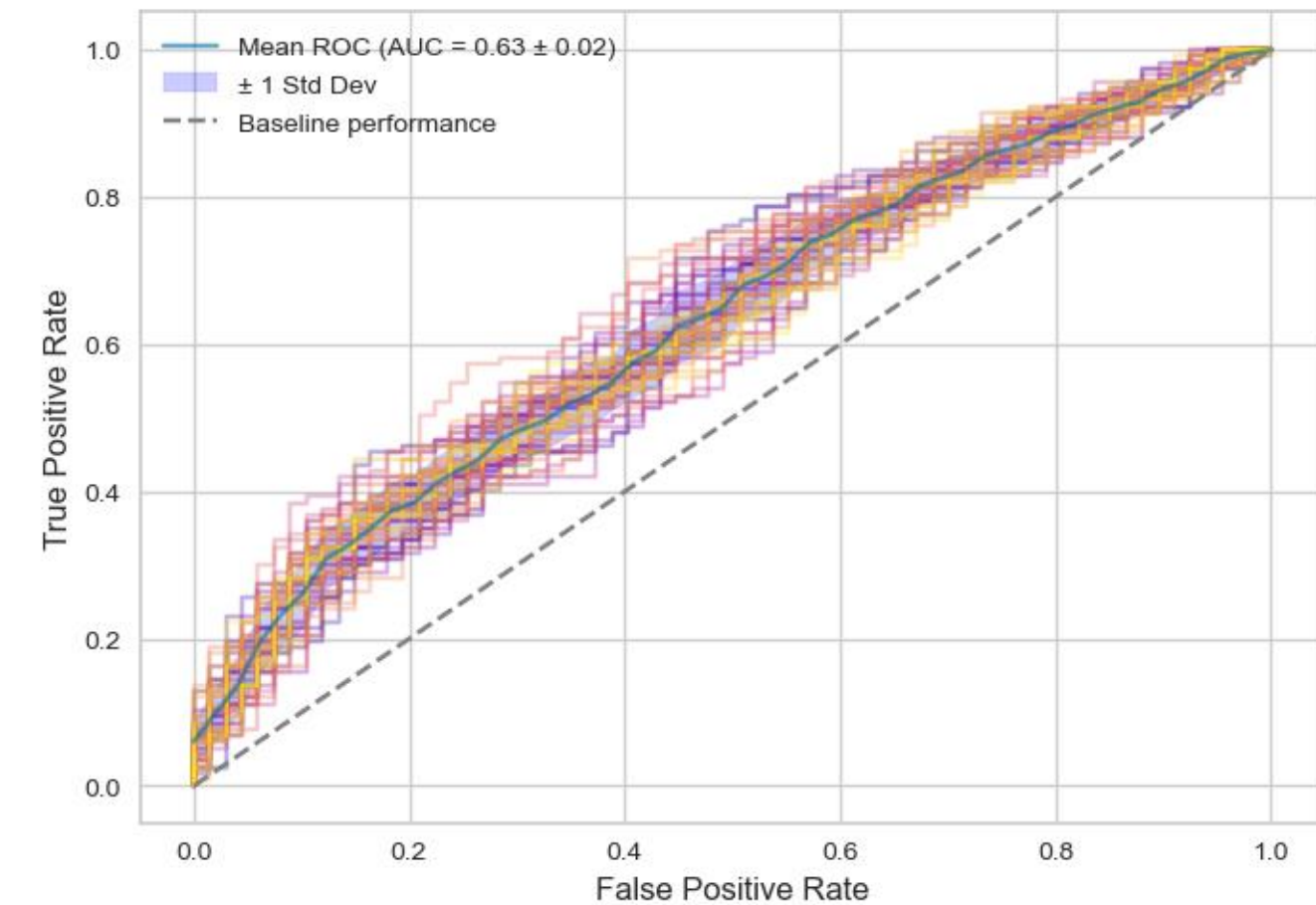
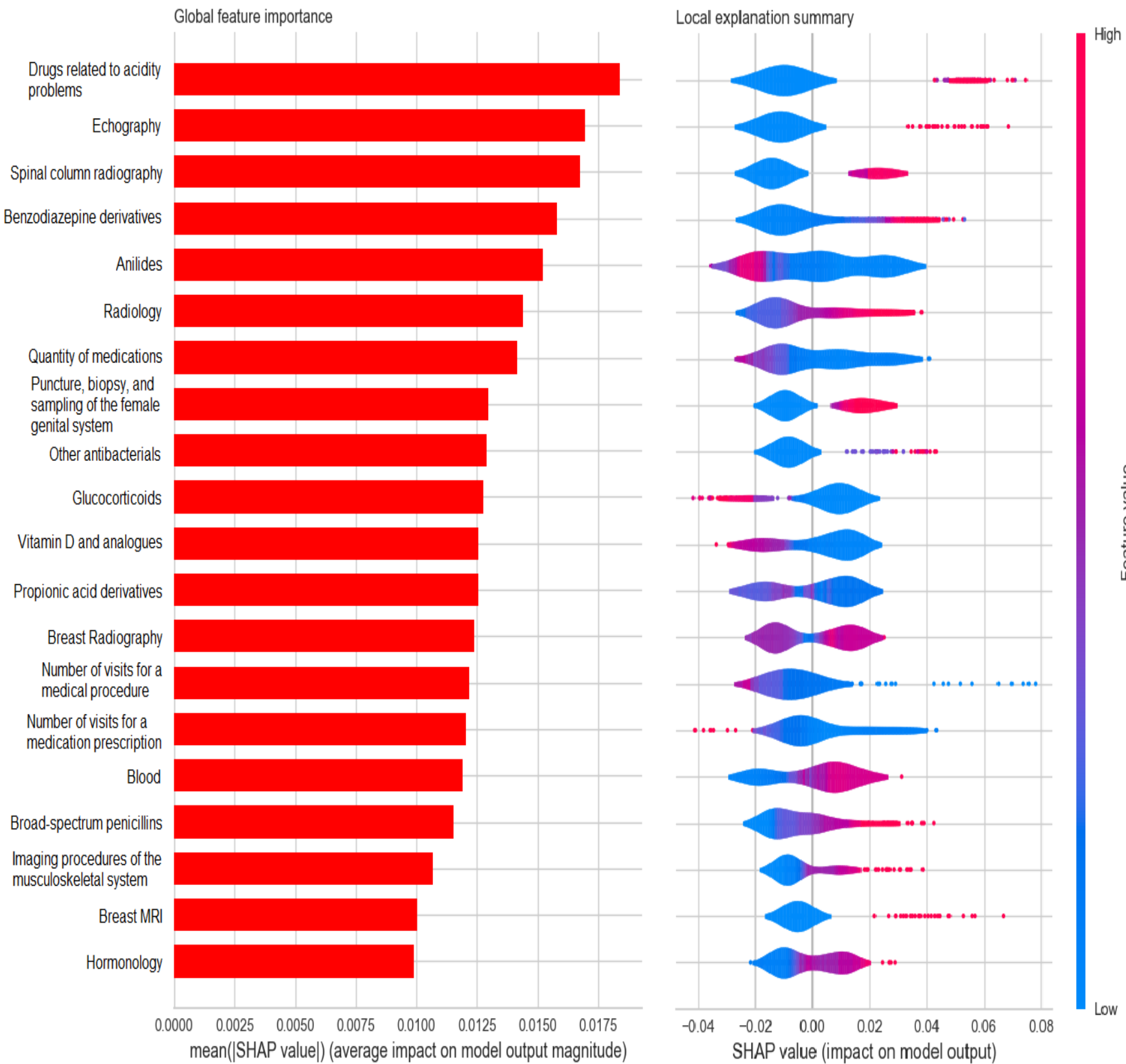


Figure. 50 Courbes ROC du modèle GB dans l'expérience 2, avec SFM et SMOTE.

- ▶ Globalement, les algorithmes de ML ont produit des résultats satisfaisants, avec ceux de l'expérience 2 se démarquant comme les plus performants.
- ▶ Les meilleurs modèles GB et MLP démontrent un meilleur équilibre à travers diverses métriques d'évaluation, notamment en termes d'AUC (aire sous la courbe ROC) mais GB se démarque en termes de Recall (identification des vrais positifs) et de F1-score (moyenne harmonique entre Précision et Recall).
- ▶ Le choix entre SMOTE et ADASYN impacte significativement les performances du modèle, SMOTE produisant généralement de meilleurs résultats.
- ▶ Les méthodes de sélection des variables ont impacté fortement les résultats.

Résultats (4/5)

Interprétation des résultats

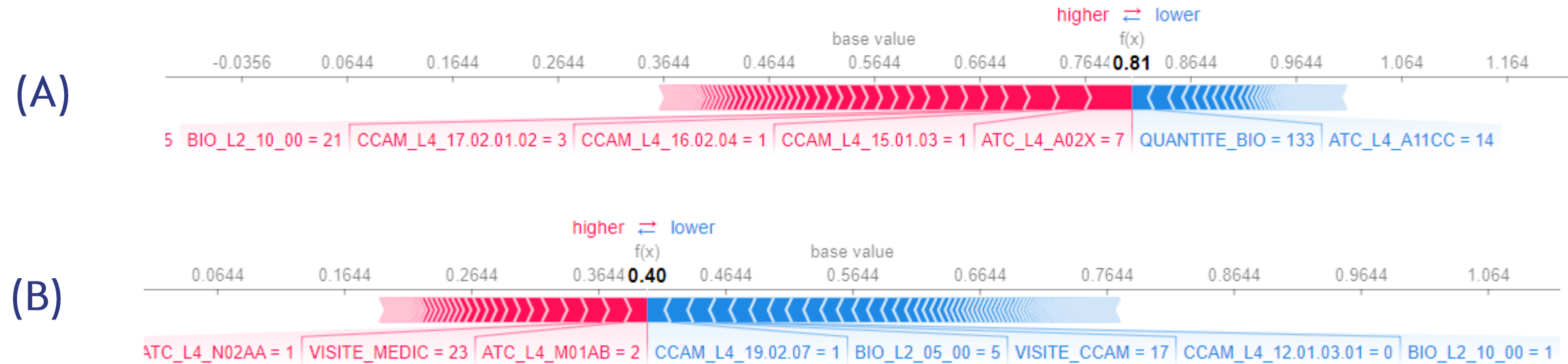


Graphique de gauche

- Illustre l'importance globale des 20 principales variables en utilisant GB avec SMOTE, SFM et l'expérience 2. : impact global sur les prédictions du modèle.
- « problèmes d'acidité », « échographie », « dérivés de benzodiazépine », « anilides » et « actes de radiologie » et « radiographie de la colonne vertébrale » présentent des valeurs SHAP élevées.
- « ponction, biopsie et échantillonnage du système génital féminin » ont un impact soit relativement moindre.

Graphique de droite

- Illustre l'explication locale des 20 principales variables : un aperçu détaillé de l'impact des différentes variables sur les prédictions du modèle.
- « problèmes d'acidité », « échographie », « radiographie de la colonne vertébrale » et « radiologie » présentent des impacts locaux positifs, soulignant leur contribution à la classe 1.
- « glucocorticoïdes », « vitamines D et analogues », et « dérivés d'acide propionique » montrent des impacts locaux positifs, indiquant leur contribution potentielle à la probabilité que l'instance soit classée comme la classe positive, et vice versa.



► Figure (A) : un individu avec une classe prédite « 0 »

- Des valeurs élevées pour « QUANTITE_BIO » et « vitamine D et analogues » (ATC_L4_A11CC) sont mises en évidence en bleu, suggérant leur capacité à augmenter la probabilité de PRC et à conduire à une classification correcte.
- Des valeurs plus élevées et un indicateur rouge pour « IRM mammaire » (CCAM_L4_16.02.04), « thérapie endocrinienne adjuvante » (BIO_L2_10_00), et « problèmes d'acidité » (ATC_L4_A02X) indiquent un impact positif sur le modèle et contribuent à une classification correcte.

► Figure (B) : un individu avec une classe prédite « 1 »

- Les valeurs de « alcaloïdes naturels de l'opium » (ATC_L4_N02AA), « dérivés de l'acide acétique » (ATC_L4_M01AB), suggèrent qu'elles augmentent la valeur SHAP pour cet individu, rendant plus probable la persistance de la PRC pour ce patient et contribuant ainsi à la classe 1.
- « actes de radiologie » (CCAM_L4_19.02.07), « hématologie » (BIO_L2_05_00) et « radiographie de la colonne vertébrale » (CCAM_L4_12.01.03.01) ont des valeurs faibles, indiquant qu'elles contribuent à la classe positive.

Analyse précieuse sur la manière dont différentes variables peuvent avoir une influence sur les prédictions du modèle : crucial pour affiner les modèles prédictifs dans le contexte de résultats psychologiques liés au cancer.

Unicité de notre ensemble de données



Aucune étude antérieure n'a été menée, à notre connaissance, pour prédire la PRC en utilisant ce type de données de remboursement.

Surconsommation médicale



Les individus du groupe PRC présentent un nombre significativement plus élevé d'actes biologiques et médicaux par rapport à ceux du groupe Non-PRC.

Résultats prometteurs



Compétitifs, soulignant leur utilité potentielle dans l'identification des patients à risque de PRC clinique.

Implications cliniques



- ❖ Le modèle offre une approche non invasive et économique pour prédire la PRC, complétant les évaluations cliniques existantes.
- ❖ Un besoin de conseils de professionnels et d'une éducation supplémentaire des femmes pour différencier les effets secondaires de l'AET (thérapie endocrinienne adjuvante) et les symptômes à long terme de ceux indiquant une rechute.
- ❖ Les médicaments liés à l'acidité et les benzodiazépines peuvent être associés à l'anxiété, et leur consommation devrait alerter sur la détresse psychologique possible du patient. Idem pour la multiplicité des actes d'imagerie et des analyses sanguines

Mesure de la PRC

- Basée sur une échelle à un seul item plutôt que sur une échelle multi-items.
- Il n'existe pas de consensus clair sur la meilleure façon de mesurer la PRC, mais la fréquence/persistance des inquiétudes est considérée comme une caractéristique clé de la PRC clinique, et fournit déjà des informations utiles.

Modèles de ML conventionnels

- Contraintes sur la taille et le type de données
- Non prise en compte du temps

Conclusions (3/3)

Approches méthodologiques

- ▶ Apprentissage semi-supervisé : exploitant des données de remboursement non étiquetées (PRC inconnue).
- ▶ Intégrer le temps pour améliorer les performances du modèle avec des réseaux de neurones : RNNs, LSTMs, GRUs, Autoencodeurs, etc.
- ▶ Capturer la nature dynamique des trajectoires de soins de santé et mieux comprendre les motifs évolutifs associés à la PRC chez les femmes atteintes de cancer du sein.
- ▶ Modèles prédictifs plus nuancés capables de saisir les changements dans les profils de risque des patients au fil du temps.
- ▶ Utilisation d'autres manières de regrouper les consommations de soins (en cours de réflexion...).

PLAN PERSONNALISÉ DE SURVEILLANCE APRES CANCER DU SEIN

Date de diagnostic :



Date d'élaboration du plan : Remis le : par : à Mr Mme :

Date :	Mois	Mois	Mois	Mois	Mois	Mois	Mois	Mois	Mois	Mois	Mois	Mois
Examens prescrits :												
Consultation Médecin Traitant												
Consultation IPC												
Mammographie												
Echographie mammaire												
IRM mammaire												
Echographie abdominale												
Thorax												
Scintigraphie osseuse												
Scanner - Type												
Ostéodensitomètre												
Evaluation cardiaque												
Bilan biologique FNS Plaquettes Bilan hépatique Bilan lipidique												
CA 15.3.												
Autres : Tep-Scan												

* Le médecin rédacteur de ce plan marque d'une croix les dates des examens qu'il prévoit pour la surveillance de sa patiente.



Remerciements !!!



Marseille, 12 avril 2024