Biomarqueurs pronostiques dans le Cancer du Colon

Lancelot Ravier, Jade Bordet, Manal belouarda, Sara Mekkaoui
Encadrantes : Ekaterina Flin, Pr. Séverine Valmary-Degano

Université Grenoble Alpes
IAB & CHUGA

18 avril 2025

Plan

Présentation de l'IAB

Institut pour l'Avancée des Biosciences

UGA, INSERM, CNRS

3 departments
19 teams
300 personnes

Signaling through Chromatin

Team « Régulations Epigénétiques »

Saadi Khochbin
(Team Leader)

18 permanent staff
- bioloogistes
- Ingénieurs informatique
- Docteurs en médecine

4 étudiants en thèse
3 étudiants en Master

EpiMed Centre de bioinformatique « EpiMed »

Organisation et outils de travail

Réunions régulières pour suivre les étapes clés du cahier des charges

Outils utilisés

- Outil de gestion collaborative : Trello
- Langage : Python, LaTeX(présentation)
- Bibliothèques utilisées : Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, Lifelines, Scipy, Statsmodels

Introduction : le cancer du colon

Le cancer du colon

En 2023 (France)

- 4^{eme} cancer le + diagnostiqué / 2^{nd} le plus mortel (hors cancer prostate / sein)
- 47 500 nouveaux cas
- Diagnostic moyen : 71-72 ans
- Taux de mortalité élevé

incertitude au stade II/IV $\rightarrow$ Besoin de nouveaux biomarqueurs pronostiques

2 types de tumeurs

- MSS (Stabilité des microsatellites) : 85% des cas
- MSI (instabilité des microsatellites) : 15% des cas, qui répondent mieux à certaines immunothérapies et ont souvent un meilleur pronostic.

Activation aberante des gènes spécifiques de tissus

Points importants

- Dérégulations génétiques dans les cellules cancéreuses
- Réactivation anormale de gènes normalement silencieux
- Méthode "ectopy" développée à l'IAB pour détecter ces gènes
- Biomarqueurs associés à un mauvais pronostic

La signature pronostique GEC

Points importants

- Méthode “ectopy” : identification d’une signature de 4 gènes (ERFE, HOXC6, LAMP, ULBP2)
- Corrélation entre nombre de gènes activés et pronostic défavorable
- Validation sur une cohorte rétrospective de 140 patients du CHUGA
- Technologies utilisées : RT-qPCR ou immunohistochimie
- Objectif : intégration du GEC en pratique clinique + analyses statistiques à réaliser

Objectifs du projet

Objectifs

Objectif 1 : Simplifier la signature

- Seulement **3 gènes détectables en pratique**
- Comparaison : **signature à 4 vs 3 gènes**
- Mesure : **impact sur la survie**

Objectif 2 : Intégrer le statut MSI/MSS

- Statut **non directement disponible**
- **Reconstruction via données moléculaires**
- **Prédiction MSI/MSS** par machine learning

Objectif final

- Vérifier si **GEC apporte une valeur indépendante**
- Tester la **fiabilité pour tous les patients**, quel que soit le statut MSI/MSS

Jeux de données et traitement

Jeux de données & pré-traitement

Données

Types de données	**TCGA-COAD**	**GSE39582**	**GSE17536**
Expression de gènes	Oui (RNA-seq*, 398)	Oui (puces**, 566)	Oui (puces**, 177)
Méthylation d'ADN	Oui	Non	Non
Mutations de gènes	Oui	Non	Non
CNA	Oui	Non	Non
Survie	Oui	Oui	Oui
Score GEC	Oui	Oui	Oui

Table – Résumé des types de données disponibles

* Technologie de séquençage de l'ARN
** Technologie de puces à ADN

Prétraitement par EpiMed $\rightarrow$ transformation *log*, normalisation

Présentation des outils d'analyse

Analyse de survie

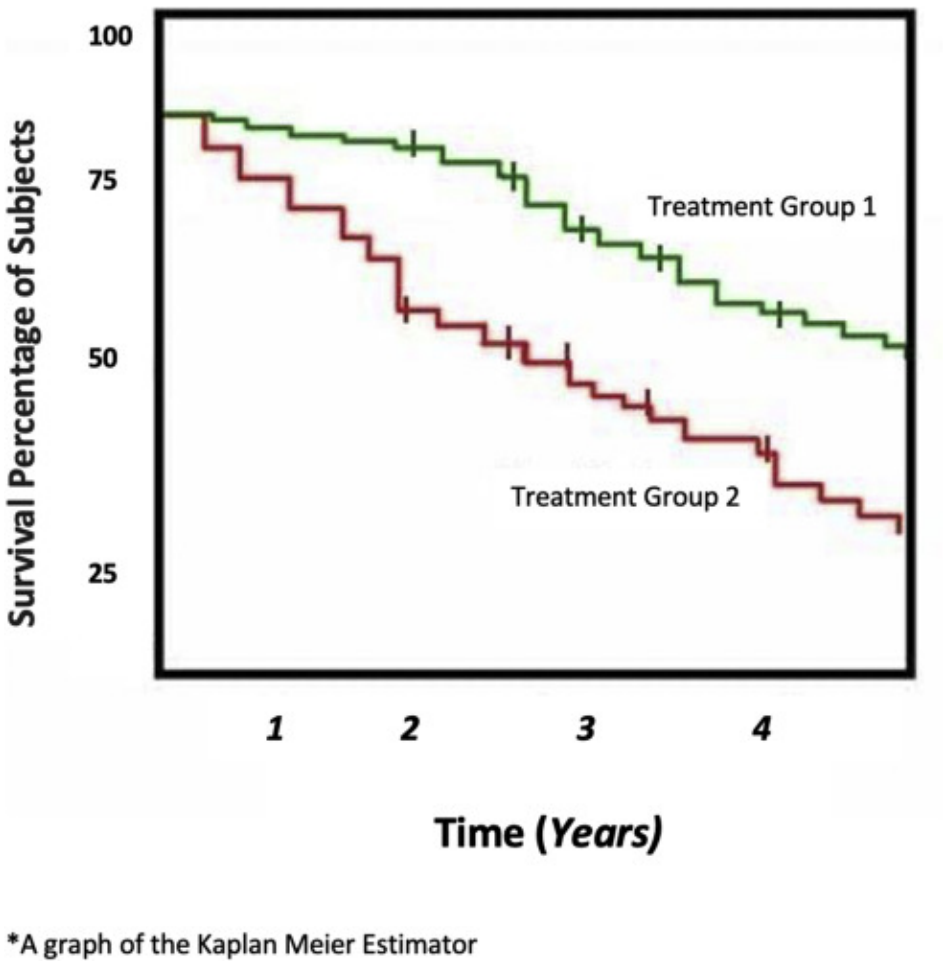

Figure – Analyse de survie avec courbes de Kaplan-Meier

Modèle de Cox

Définition de la regression de cox

$h(t \mid Z^i) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k Z_k^i\right) = h_0(t) e^{\theta^T Z}$

- $h(t \mid Z^i)$ est le taux de risque instantané pour l'individu i,
- $h_0(t)$ est la fonction de risque de base, indépendante des covariables,
- Z^i est le vecteur des covariables pour l'individu i,
- $\theta = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients de régression inconnus.

Hazard Ratio

$HR = e^{\beta_j}, \quad \forall j \in \{1, \ldots, p\}$

- $HR > 1$: la variable augmente le risque
- $HR < 1$: la variable protège
- $HR = 1$: pas d'effet

Test du log-rank

Statistique de test

$U = \sum_{i=1}^{k} d_{B,i} - \frac{R_{B,i}}{R_{A,i}+R_{B,i}}(d_{A,i} + d_{B,i})$

- k : le nombre total d'instants de décès observés (événements),
- $d_{A,i}, d_{B,i}$: le nombre d'événements (décès) aux temps t_i dans les groupes A et B,
- $R_{A,i}, R_{B,i}$: le nombre de patients à risque juste avant t_i dans les groupes A et B.

Règle de décision

Rejet de H_0 si : $\mathbb{P}_{H_0}(T_n > t_{n,\text{obs}}) \leq \alpha$
Sous H_0 : différences entre les groupes dues au hasard

Résultats

Analyse de survie : Comparaison 4-GEC vs 3-GEC

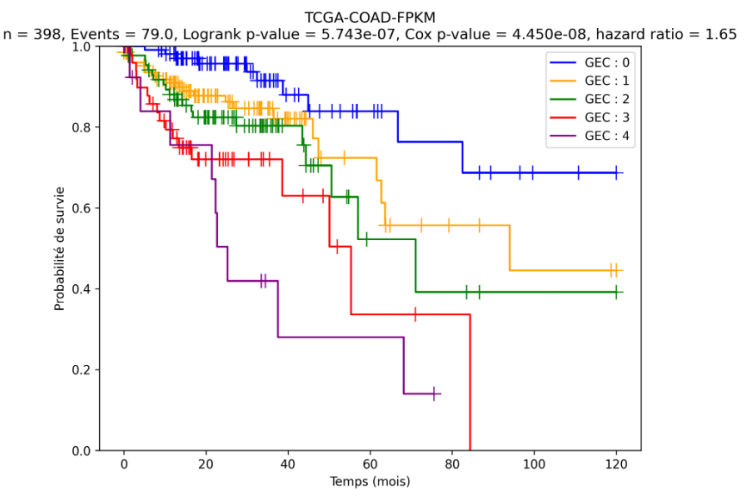

Figure – TCGA-COAD – 4-GEC

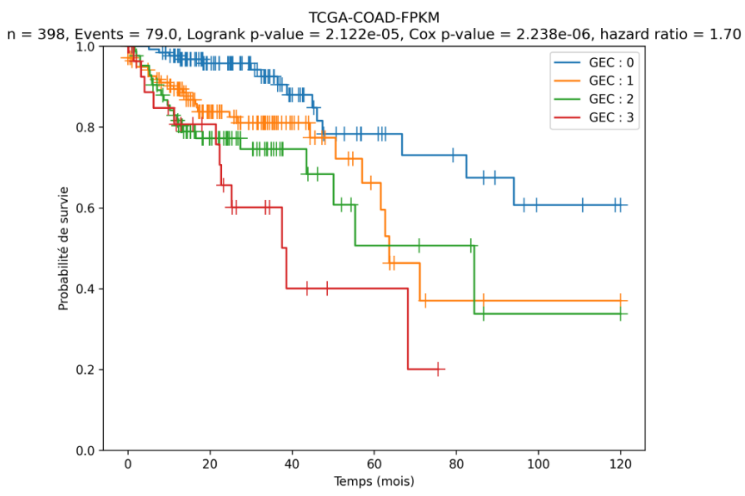

Figure – TCGA-COAD – 3-GEC

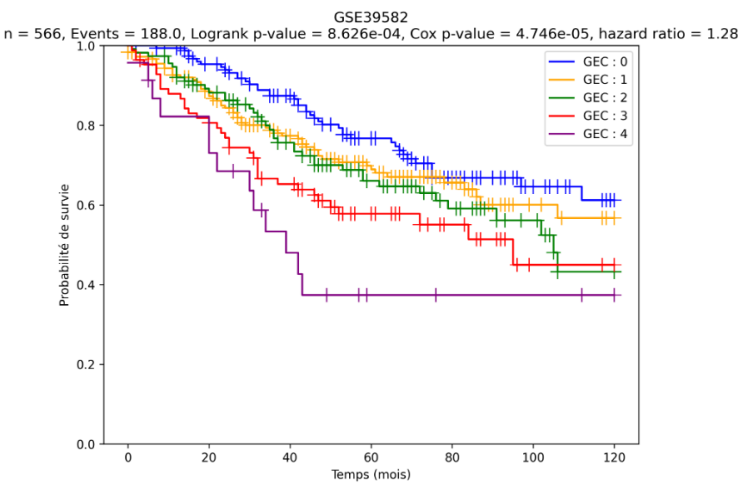

Figure – GSE17536 – 4-GEC

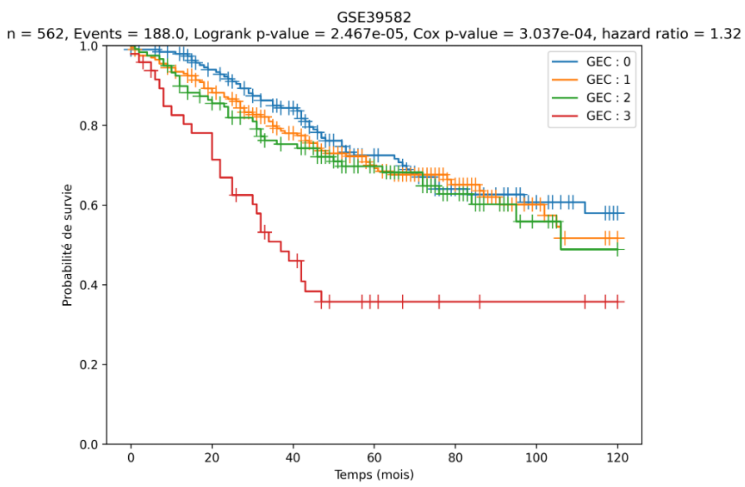

Figure – GSE17536 – 3-GEC

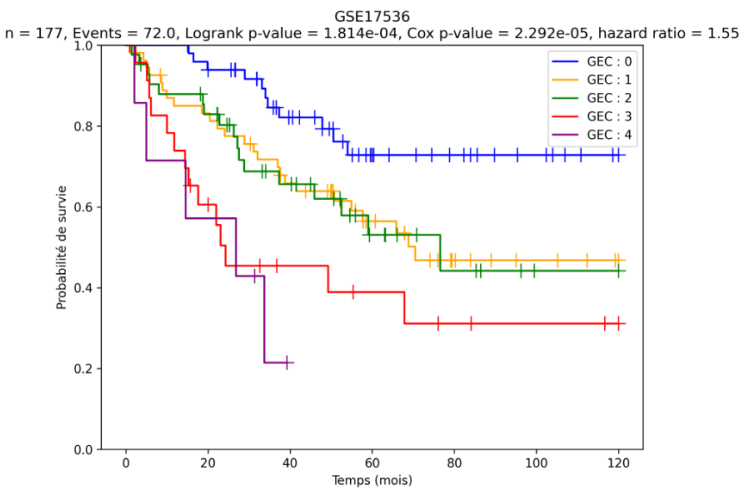

Figure – GSE39582 – 4-GEC

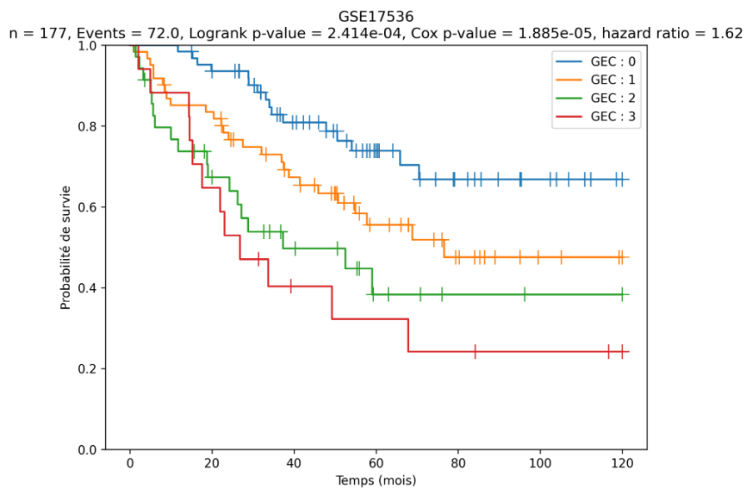

Figure – GSE39582 – 3-GEC

Conclusion

- La signature 4-GEC reste la plus précise.
- La version 3 gènes est une alternative plus simple avec un bon pouvoir pronostique.
- Compatible avec l'immunohistochimie (méthode courante, peu coûteuse, accessible).
- Favorise une intégration rapide en pratique clinique hospitalière.

Détermination du statut MSI/MSS (dataset TCGA-COAD)

Détermination du statut MSI/MSS (dataset TCGA-COAD)

Conditions MSI

- Méthylation du gène MLH1 : methylation moyenne de la région promotrice + faible expression.
- Mutations des gènes MLH1, MSH2, MSH6, PMS2 : mutation LOF (Loss Of Function - Perte de fonction).
- Perte de gènes par altération du nombre de copies (CNA) : perte d'une région entière de l'ADN contenant l'un de ces gènes.

MSI : au moins l'une des 3 conditions

Identification MSI-MSS - Données de Methylation

Données

- $n = 398$
- plusieurs régions promotrices methylées $\rightarrow$ moyenne calculée
- MSI (bleu) : expression < 1.5 & methylation moyenne > 0.3 ($n = 34$)

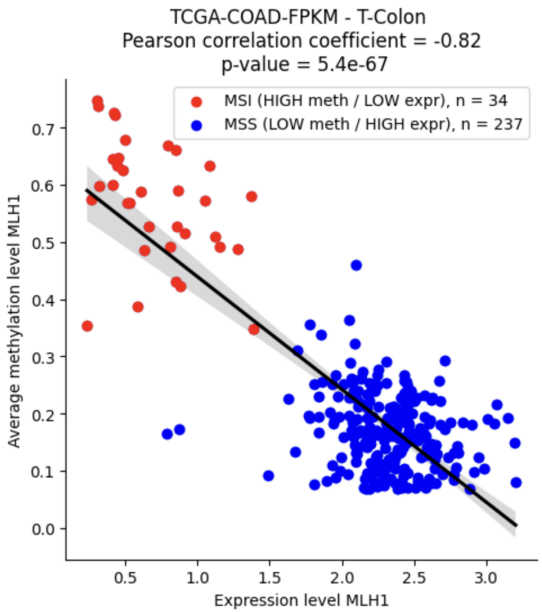

Figure – Donneés de methylation moyenne en fonction des données d'expression pour chaque sample

Identification MSI-MSS - Données de Mutation

Analyse des mutations

- 362/398 échantillons analysables
- 20 avec mutation $\rightarrow$ MSI
- 342 sans mutation $\rightarrow$ besoin d'autres données

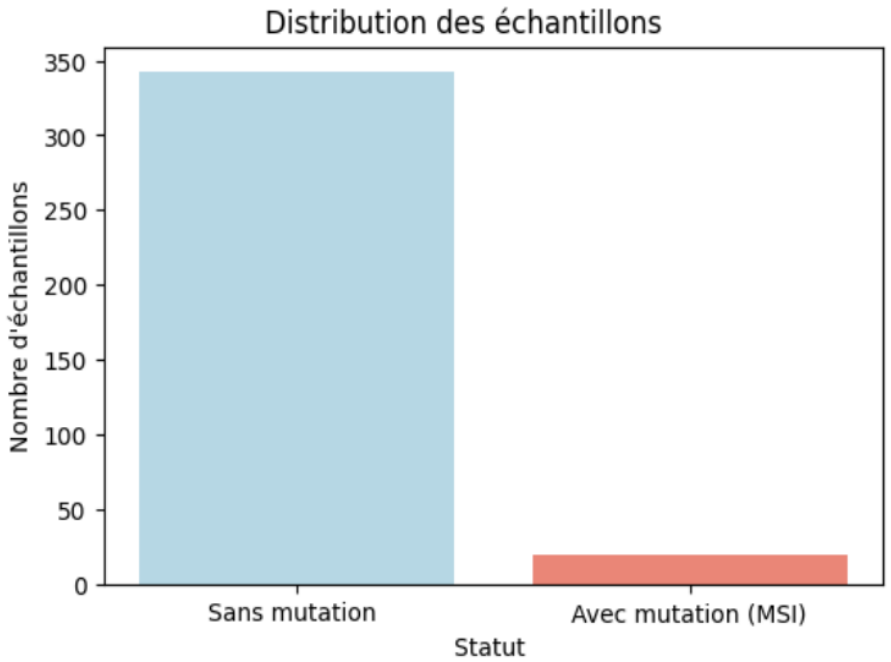

Figure – Donneés de methylation moyenne en fonction des données d'expression pour chaque sample

Identification MSI-MSS - Données de CNA

Analyse des CNA

- 395 échantillons analysables
- 12 avec alteration du nombre de copies → MSI
- 383 sans CNA → besoin d'autres données

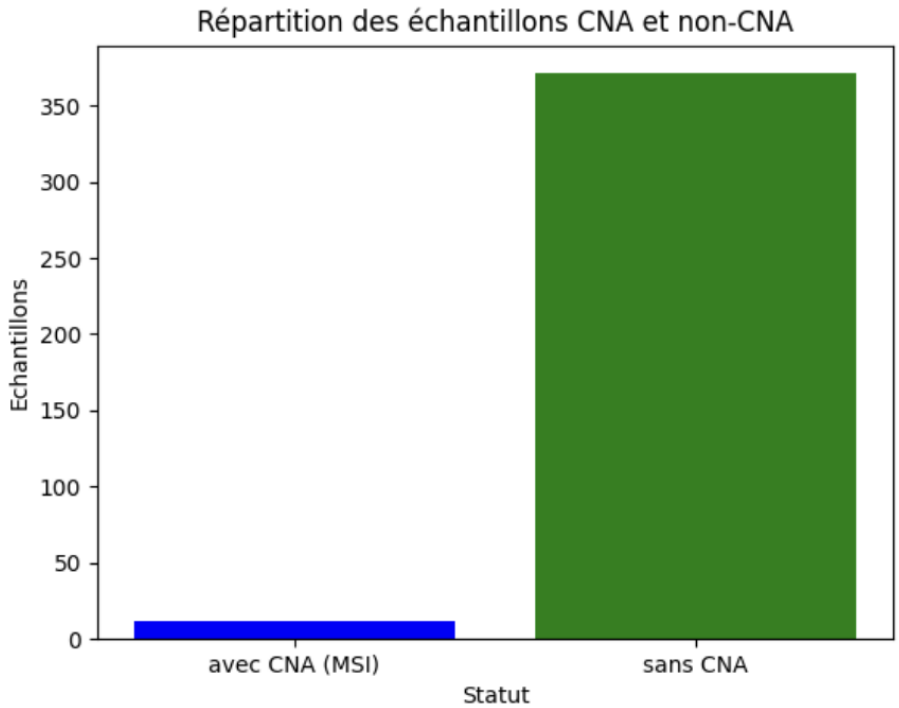

Figure – Donneés de CNA des samples TCGA-COAD

Prédiction MSI/MSS par machine learning

Résultats finaux - statut MSI/MSS

Statut	Nombre d'échantillons
MSI	56
MSS	212
Inconnu	130
Total	398

Table – Répartition des échantillons selon le statut MSI/MSS

- 217 gènes sélectionnés via la littérature et analyse différentielle
- Nettoyage des données (NA, normalisation, codage binaire de la cible)
- 268 échantillons (MSI + MSS), 217 gènes

Résultats des modèles de prédiction

Modèles de machine learning

Modèle	Accuracy	Precision	F1-score
XGBoost_Tuned	0.925373	0.973684	0.787234
XGBoost	0.910448	0.833333	0.769231
Extra Trees	0.899254	0.891892	0.709677
Neural Network	0.895522	0.780000	0.735849
Random Forest	0.888060	0.809524	0.693878
SVM	0.884328	0.735849	0.715596
Decision Tree	0.873134	0.683333	0.706897
Lasso reg.	0.869403	0.677966	0.695652
Linear reg.	0.828358	0.562500	0.661765
Ridge reg.	0.828358	0.562500	0.661765
Naïve Bayes	0.828358	0.564103	0.656716
Elastic Net reg.	0.820896	0.546512	0.661972

Figure – Modèles de machine learning et métriques principales

Impact pronostique de la signature 4-GEC et résultats pour panel 3-GEC

Résultats finaux

- Un modèle performant pour prédire MSI/MSS dans TCGA à partir de l'expression génique : XGBoost Tuné
- Limite : généralisation difficile à d'autres datasets (manque d'annotations + hétérogénéité technologique)

Analyse de la signature GEC dans les groupes MSI/MSS

Objectif

- Evaluer les scores GEC des groupes MSI/MSS
- Prédiction de la survie globale

Modèles utilisés dans la prédiction du statut MSI/MSS

- XGBoost
- SVM

Métriques

- Logrank p-value : teste la différence entre les courbes de survie (GEC 0 v.s. GEC 1-4)
- Hazard Ratio (HR) : Mesure le risque relatif associé au GEC
- C-Index : Mesure la capacité discriminatoire du modèle

Résultats pour MSI/MSS

Résultats MSI

Modèle	Logrank p	Cox p	HR (GEC)	C-index
XGBoost_Tuned	0.0456	0.0169	1.75	0.660
SVM	0.0250	0.00411	1.88	0.683
RF	0.0338	0.00949	1.79	0.670

Résultats MSS

Modèle	Logrank p	Cox p	HR (GEC)	C-index
XGBoost_Tuned	0.0456	0.0169	1.75	0.660
SVM	0.0250	0.00411	1.88	0.683
RF	0.0338	0.00949	1.79	0.670
Extra_trees	0.0314	0.00976	1.79	0.670
XGBoost	0.0330	0.0115	1.76	0.667
MLP	0.0452	0.0296	1.68	0.658

Analyse de la signature GEC dans les groupes MSI/MSS

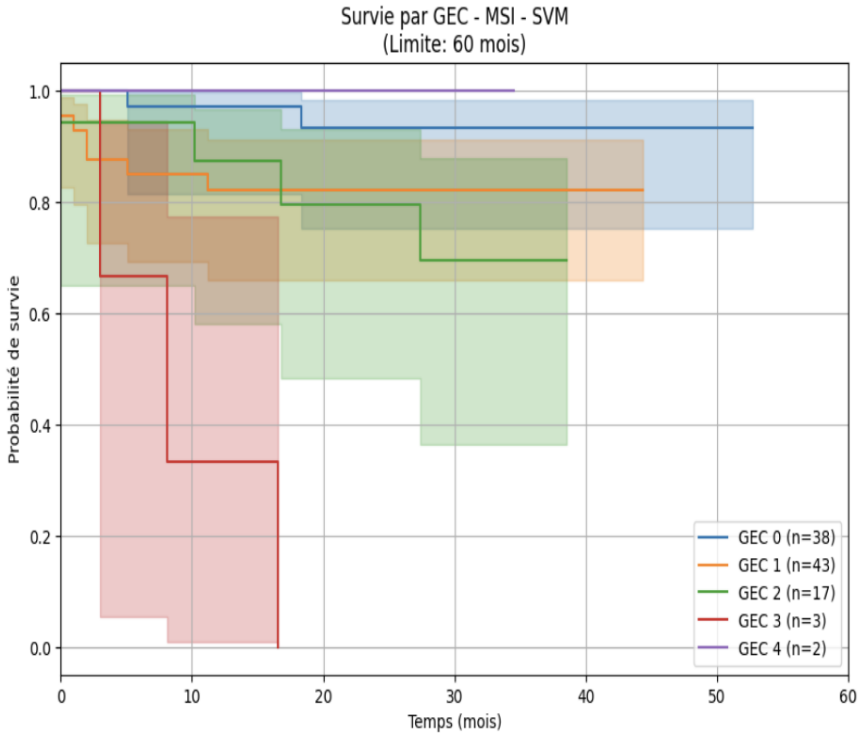

Figure – Survie par GEC - MSS

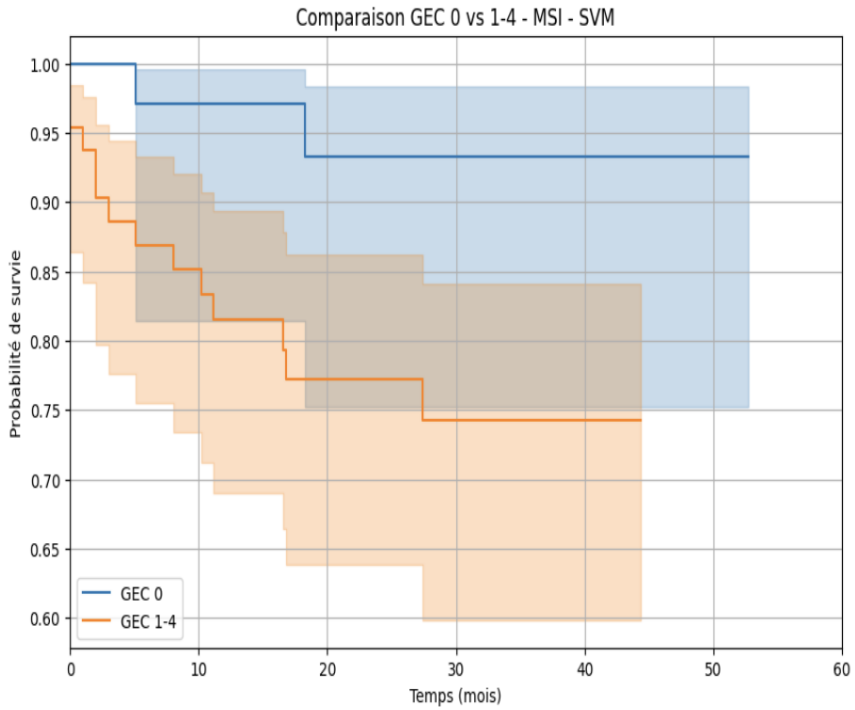

Figure – Comparaison GEC 0 v.s. 1-4

Résultats

Analyse multivariée ($n = 268$)

- GEC $\geq 3 \rightarrow HR = 3.80$
- GEC $\geq 4 \rightarrow HR = 4.42$
- $p < 5\%$ dans les deux cas
- Statut MSI/MSS non significatif

- Dans cette cohorte, le GEC est le seul marqueur prédictif solide
- Mème une simple comparaison de survie entre les patients MSI/MSS, sans prendre en compte le GEC, ne montre aucune différence significative

Impact sociétal et environnemental

Points d'impact et consommation de CO2

- Impact postitif médical et sociétal
- Enjeux environnementaux liés au numérique
- Estimation de l'empreinte carbone : $\sim 35 kgCO_2$
- Développement de 3-GEC : immunohistochimique plus écologique que la RT-qPCR

Conclusion

Conclusion

Signature GEC

Signature GEC $\rightarrow$ significatif dans la prediction de survie des patients T-colon

Score GEC

Score GEC indépendant du statut MSI-MSS $\rightarrow$ information non-redondante

Modèle de prédiction du statut MSI/MSS

Prédiction du statut MSI/MSS à partir des donneés d'expression (ML) $\rightarrow$ utilité pour des données sans annotation

3-GEC et impact ecologique

Version simplifiée 3-GEC (out ERFE) significatif $\rightarrow$ test applicable en milieu hospitalier sans production d'anticorps supplémentaire