Evaluation de différentes variantes du modèle de Cox pour le pronostic de patients atteints de cancer à partir de données publiques de séquençage et cliniques

Rémy Jardillier

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Rémy Jardillier

Thèse dirigée par **Laurent GUYON,** Docteur, CEA, et co-encadrée par **Florent CHATELAIN,** Maître de conférences, Grenoble-INP

préparée au sein du **Laboratoire BCI - Biologie du Cancer et de l'Infection**
dans **l'École Doctorale Ingénierie pour la Santé la Cognition et l'Environnement**

Évaluation de différentes variantes du modèle de Cox pour le pronostic de patients atteints de cancer à partir de données publiques de séquençage et cliniques

Thèse soutenue publiquement le **1 décembre 2020**,
devant le jury composé de :

Mme Anne-Laure BOULESTEIX
PROFESSEUR, Université de Munich, Rapporteure
M. Christophe AMBROISE
PROFESSEUR DES UNIVERSITES, Université d'Évry Val d'Essonne, Rapporteur
Mme Adeline LECLERCQ-SAMSON
PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes, Présidente
M. Vivian VIALLON
MAITRE DE CONFERENCE, Centre International de Recherche sur le Cancer (IARC), Examinateur
M. Laurent GUYON
DOCTEUR EN SCIENCES, CEA, Directeur de thèse

et du membre invité :

M. Florent CHATELAIN
MAITRE DE CONFERENCES, Grenoble-INP, Co-encadrant

Remerciements

Mes premiers mots vont vers l'équipe IMAC (Odile, Nadia C., Nadia A., Isabelle, Josiane, Aude, Caroline, Soha, Inrinka, Constance, Cathy, Laurent, Christophe, Fred, Nicolas, Claude), qui m'a accueilli à bras ouverts pour ces trois années de thèse. Étant donné les allergies de certains pour les formules mathématiques, j'ai fait de mon mieux pour vulgariser mon sujet. Merci pour l'intérêt que vous lui avez porté, ainsi que pour vos questions, vos suggestions, et votre enthousiasme. J'ai été impressionné par la passion et le cœur que vous mettez à l'ouvrage, et je suis convaincu de l'intérêt de vos recherches pour la science et le bien-être de tous. Malgré les contraintes de plus en plus fortes qui sont exercées sur le métier de chercheur, votre passion demeure la plus forte, et j'espère que cela continuera ! Ensuite, je remercie Claire, Irinka, Quentin et Clément, avec qui j'ai partagé le bureau, pour votre gentillesse et votre bienveillance. J'ai été heureux de vous rencontrer, et votre présence au quotidien a participé à mon équilibre et à ma sérénité. Vous trouverez en Cathy une successeure de grandes valeurs, dont les fleurs rouges, les qualités humaines et la bonne humeur viendront égaillés le bureau ! Enfin, je me suis largement inspiré des diapositives de Soha pour ma soutenance, et le rendu aurait certainement été beaucoup moins clair sans cela. Comme promis, je te remercie ici !

Laurent et Florent, je vous remercie pour vos conseils scientifiques, pour votre humanité dans les moments difficiles, et pour avoir réussi à me recadrer dans ceux de doute. L'initiation d'une nouvelle thématique de recherche est un projet ambitieux, qui nécessite une montée en compétence importante, un ajustement permanent des questions que l'on se pose, et un positionnement clair par rapport à la littérature. Vous avez su gérer les hauts et les bas, en ajustant les curseurs de l'encadrement, de la liberté, et de la confiance. J'espère que vous continuerez à pousser ces modèles de survie vers plus de performances et de potentielles utilités pratiques.

Les moments difficiles, comme ceux de joie, sont décisifs dans la vie d'un Homme. La présence des personnes que l'on aime est alors essentielle dans l'accompagnement des transformations qui s'opèrent en nous. C'est une grande source de bonheur et de soulagement de savoir qu'il est possible de compter sur sa famille et ses amis dans ces moments particuliers. Je souhaite donc remercier, même si le mot n'est pas assez fort, toute ma famille (Papa, Stéphanie, Myriam, Sylvie, Fabrice, Florence, Marie-Pierre, Pierre, mes grands-parents - Arlette, Paul, Michèle et Jean, mon arrière-grand-mère - Andrée, Thaïs, Zélie, Milo, Guillaume, Caroline, Pierre-Antoine, Claire), ainsi que mes amis proches (Géraud, Anthony, Paul, Luis, Benjamin, Vincent, Antoine, Jorge, Romain, Hugo, Maximilien, Florent, Thibaut, Arnaud, Théo et j'en passe...). Je vous exprime ici toute ma reconnaissance et mon amour.

Enfin, mes derniers remerciements vont vers Maud. Tu as su me soutenir, à bout de bras parfois, avec force, détermination, et de manière inconditionnelle. Ta lumière, ton courage et ta douceur resteront gravés pour toujours, et j'espère qu'ils continueront à

égayer ma vie jusqu'à sa fin.

Je voudrais terminer par dédier cette thèse à ma mère. Après un début poussif et des moments de doute, c'est pour elle, pour la rendre fière, que j'ai essayé de mener à bien ce projet, du mieux que j'ai pu.

Table des matières

Liste des figures vi

Liste des tableaux viii

Contributions ix

1 Introduction 1

1.1 Cancer, recherche, et prédiction de la survie . . . 3
1.2 Les molécules d'ARN . . . 5
1.3 Le séquençage : mesure du niveau d'expression des gènes . . . 8
1.4 Normalisation des données de comptage RNA-seq . . . 12
1.5 La base de données TCGA . . . 15
1.6 Modélisation des données de survie . . . 17
1.7 Métriques d'évaluation de la qualité de prédiction . . . 21
1.8 La « malédiction de la grande dimension » . . . 28
1.9 Conclusions . . . 33
1.10 Références . . . 35

2 Comparaison et évaluation des méthodes de pénalisation du modèle de Cox avec les données mRNA-seq 45

2.1 État de l'art et objectifs du chapitre . . . 47
2.2 Comparaison des capacités de prédiction des méthodes de pénalisation du modèle de Cox sur données réelles . . . 48
2.3 Combinaison des données cliniques et mRNA-seq pour prédire la survie . . . 56
2.4 Procédure de simulation des données de survie et d'évaluation des pénalisations . . . 60
2.5 Capacités de prédiction sur données simulées . . . 68
2.6 Capacités de sélection des méthodes de pénalisation du modèle de Cox . . . 74
2.7 Influence des performances de sélection sur la prédiction . . . 80
2.8 Conclusions . . . 84
2.9 Références . . . 85

3 Pré-filtrage univarié des données d'expression génétiques et prédiction avec le modèle de Cox multivarié **91**
3.1 Pré-filtrage univarié des données d'expression génétique 92
3.2 Impact du pré-filtrage sur les capacités de prédiction du modèle de Cox pénalisé 96
3.3 Comparaison des méthodes de pénalisation du modèle de Cox après pré-filtrage 102
3.4 Comparaison du pré-filtrage bi-dimensionnel à l'algorithme *Iterative Sure Independance Screening* 102
3.5 Conclusions et perspectives 106
3.6 Références 107

4 Impact de la profondeur de séquençage des données miRNA-seq sur la prédiction **111**
4.1 Contexte de l'étude 112
4.2 Choix des cancers étudiés 114
4.3 Comparaison des prédictions obtenues avec les ARNm et les miARN 118
4.4 Prédiction de la survie avec les variables cliniques et les données miRNA-seq 121
4.5 Dégradation des données de séquençage 123
4.6 Impact de la taille des banques et du nombre de patients sur la prédiction . 125
4.7 Cause de la dégradation de la qualité de prédiction 129
4.8 Conclusions 133
4.9 Références 134

5 Conclusions et perspectives **139**
5.1 Conclusions générales 140
5.2 Résultats préliminaires et perspectives 141
5.3 Conclusions pratiques 150
5.4 Références 151

A Figures et Tableaux Annexes **I**
A.1 Chapitre 2 I
A.2 Chapitre 3 XII
A.3 Chapitre 4 XV

B Liste des acronymes **XVII**

C Glossaire **XIX**

Résumé

Liste des figures

1.1 Transcription de l'ADN en ARN (synthèse des ARN), et traduction de l'ARN en protéine (synthèse des protéines). 7
1.2 Différentes étapes du séquençage haut-débit (RNA-seq) 11
1.3 Exemple de biais induit par la normalisation CPM 14
1.4 Courbes de survie obtenue à partir des données TCGA pour quatre cancers. 19
1.5 Fonction de risque pour différents patients. 22
1.6 Schéma récapitulatif de l'évaluation des modèles de prédiction. 24
1.7 Score de Brier en fonction du temps 27
1.8 Calcul du poids de la pénalité elastic net qui minimise la déviance globale. . 31
1.9 Déviance globale et nombre de gènes sélectionnés pour différentes valeurs de α dans la pénalisation elastic net pour les données mRNA-seq de KIRC. . 32

2.1 C-index médians obtenus avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés. 49
2.2 C-index médians en fonction des IBS médians calculés avec la pénalisation elastic net pour l'ensemble des 26 cancers de TCGA étudiés. 51
2.3 C-index, p-valeurs du modèle de Cox univarié, et IBS obtenus avec les données mRNA-seq. 54
2.4 Nombre de gènes sélectionnés, temps de calcul médian (s), et nombre d'erreurs des méthodes de pénalisation obtenus avec les données mRNA-seq. . 55
2.5 C-index obtenus avec les données cliniques, mRNA-seq, et en mixant les deux types de données. 60
2.6 Courbes de Kaplan-Meier des données de survie obtenues avec différents paramètres de simulation et sur données réelles pour LGG. 65
2.7 Histogrammes des indices pronostiques oracles obtenus avec différents paramètres de simulation et sur données réelles pour LGG. 66
2.8 Schéma récapitulatif des procédures de simulation et d'évaluation du modèle de Cox pénalisé sur données simulées. 67
2.9 C-index médians oracles et estimés, et différences médianes entre les C-index oracles et estimés pour différents paramètres de simulation pour LGG et la pénalisation elastic net - données simulées, elastic net. 69

2.10 Médiane des p-valeurs du modèle de Cox univarié (-log10) oracles en fonction du nombre de patients pour les 26 cancers de TCGA étudiés - données simulées. 69
2.11 C-index médians obtenus avec elastic net et ridge pour différents paramètres de simulations contenant la même quantité d'information pour les 26 cancers de TCGA étudiés - données simulées. 70
2.12 Métrique oracle en fonction de la métrique estimée par le modèle de Cox avec pénalisation elastic net pour LGG - données simulées. 71
2.13 Différences médianes entre les C-index oracles et estimés en fonction du nombre de patients et du taux de censure pour l'ensemble des 26 cancers étudiés. 73
2.14 Stabilité des gènes sélectionnés. 75
2.15 Sensibilités et taux de fausses découvertes (FDP) obtenus avec λ_{min} et λ_{1se} pour LGG et la pénalisation elastic net - données simulées. 76
2.16 Sensibilité, FDP, et ratio sensibilité sur FDP pour les 26 cancers de TCGA étudiés obtenus avec la pénalité elastic net - données simulées. 77
2.17 Comparaison des sensibilités et des FDP obtenus avec les pénalisations lasso, elastic net et adaptive elastic net pour les 26 cancers de TCGA étudiés - données simulées. 78
2.18 Sensibilité en fonction du taux de fausses découvertes sur les 26 cancers de TCGA étudiés et la pénalisation elastic net - données simulées. 79
2.19 Sensibilité, FDP, et nombre de gènes sélectionnés en fonction du nombre de patients dans le jeu d'apprentissage sur les 26 cancers de TCGA étudiés et la pénalisation elastic net - données simulées. 80
2.20 C-index médians obtenus si la vérité terrain était connue, et différences médianes avec les C-index obtenus avec l'ensemble des gènes, et les C-index oracles pour elastic net et LGG - données simulées. 82

3.1 Relation entre l'écart interquartile et la médiane dans les données de comptage RNA-seq, CPM, log2-CPM, et VST pour KIRC. 94
3.2 C-index médians obtenus pour différents seuils de pré-filtrage pour LGG. . 98
3.3 Seuils optimaux pour la pénalisation elastic net, le C-index, et l'ensemble des 16 cancers. 99
3.4 C-index obtenus avant et après le pré-filtrage maximisant le C-index pour les quatre méthodes de pénalisation pour LGG. 100
3.5 C-index obtenus après prè-filtrage pour les quatre méthodes de pénalisation et 16 cancers de TCGA. 102
3.6 Valeurs absolues des coefficients β en fonction des p-valeurs d'un test du rapport de vraisemblance corrigées par la méthode de Benjamini-Hochberg calculés à partir d'un modèle de Cox univarié pour LGG. 104

3.7 C-index obtenus sans pré-filtrage, avec le pré-filtrage bi-dimensionnel et la pénalisation elastic net, et avec l'algorithme *Sure Independance Screening* pour les 16 cancers de TCGA étudiés. 105

4.1 Nombre de publications référencées dans PubMed traitant des miARN et du cancer entre 2003 et 2019. 112
4.2 *Boxplot* des C-index pour 25 cancers de TCGA (miARN) 116
4.3 Taille des banques pour 25 cancers de TCGA (miARN et mARN) 118
4.4 C-index obtenus avec les données RNA-seq pour les ARNm et les miARN pour 25 cancers de TCGA 120
4.5 C-index obtenus avec les données cliniques, miRNA-seq, et en combinant les deux types de données. 123
4.6 Procédure d'évaluation de l'impact de la dégradation des données de comptage RNA-seq et du nombre de patients sur la prédiction. 125
4.7 Impact de taille des banques et du nombre de patients du jeu d'apprentissage sur le C-index pour KIRC 127
4.8 Impact de la dégradation de la taille des banques des données miRNA-seq pour KIRC sur le nombre de gènes détectés et leur niveau d'expression. . . . 130
4.9 Evaluation des deux hypothèses pour expliquer la diminution des capacités de prédiction induite par la dégradation des données miRNA-seq. 132

5.1 C-index obtenus avec elastic net avec uniquement le niveau d'expression des gènes (bleu), et avec l'ajout des termes d'interactions (jaune), et proportions de termes quadratiques sélectionnés.. 143
5.2 C-index obtenus après l'ajout de nouvelles variables cliniques pour BRCA et KIRC. 144
5.3 C-index obtenus avec les données mRNA-seq, de proportions de types cellulaires (*i.e.* CIBERSORT et xCell), et de pureté de la tumeur pour 18 cancers de TCGA. 146
5.4 C-index obtenus avec les variables cliniques et les indices pronostiques des données mRNA-seq (bleu), et avec les variables cliniques et les indices pronostiques des données mRNA-seq, xCell, et de pureté de la tumeur (jaune). 147
5.5 *Heatmap* du niveau d'expression standardisé des gènes pour BLCA, et corrélation avec la pureté de la tumeur. 148
5.6 C-index obtenus avec le modèle de Cox pénalisé (ridge) et un algorithme de forêts aléatoires sur les données miRNA-seq. 149

A.1 P-valeurs médianes du modèle de Cox univarié ($-log_{10}$) obtenues avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés. I

A.2 IBS obtenus avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés. II
A.3 C-index, p-valeurs, et IBS obtenus avec les données mRNA-seq. III
A.4 Nombre de gènes sélectionnés, temps de calcul médian (s), et nombre d'erreurs des méthodes de pénalisation obtenus avec les données mRNA-seq. . IV
A.5 C-index obtenus avec les données cliniques, mRNA-seq, et en mixant les deux types de données. V
A.6 Médianes des p-valeurs du modèle de Cox univarié (-log10) et des IBS obtenus pour différents paramètres de simulations contenant la même quantité d'information pour les 26 cancers de TCGA étudiés - données simulées. . . VI
A.7 Différences médianes entre les p-valeurs oracles (-log10) et estimées en fonction du nombre de patients et du taux de censure pour l'ensemble des 26 cancers étudiés. VI
A.8 Différences médianes entre les IBS estimés et oracles en fonction du nombre de patients et du taux de censure pour l'ensemble des 26 cancers étudiés. . IX
A.9 Stabilité des gènes sélectionnés. IX
A.10 P-valeurs et IBS obtenus après prè-filtrage pour les quatre méthodes de pénalisation et 16 cancers de TCGA. XIII
A.11 Seuils optimaux pour la pénalisation elastic net, l'ensemble des 26 cancers, et la p-valeur du modèle de Cox univarié et l'IBS. XIII
A.12 P-valeurs du modèle de Cox univarié (-log10) et IBS obtenus sans pré-filtrage, avec le pré-filtrage bi-dimensionnel et la pénalisation elastic net, et avec l'algorithme ISIS pour 16 cancers de TCGA. XIV
A.13 Taille des librairies pour 25 cancers de TCGA (mARN) XV
A.14 Impact de la dégradation de la taille des banques des données mRNA-seq sur le nombre de gènes détectés et leur niveau d'expression pour KIRC. . . . XV
A.15 Cause de la diminution des capacités de prédiction induite par la dégradation des données mRNA-seq. XVI

Liste des tableaux

1.1 Cancers retenus et données de survie utilisées. 16
1.2 Tableau récapitulatif des données de survie et mRNA-seq pour les 26 cancers de TCGA retenus. 33

2.1 Corrélations entre les trois métriques d'évaluation des prédictions. 51
2.2 Variables cliniques présentes pour chacun les 26 cancers étudiés. 58
2.3 Médianes des C-index oracles, des p-valeurs du modèle de Cox univarié (-log10) oracles, et des IBS oracles pour les 26 cancers de TCGA étudiés - données simulées. 68
2.4 Médianes des différences entre C-index oracles et estimés par le modèle de Cox pénalisé pour les 26 cancers de TCGA étudiés - données simulées. . . . 72
2.5 Comparaison des C-index oracles, estimés, et estimés avec uniquement les gènes de la vérité terrain pour les 26 cancers de TCGA - données simulées. . 83

3.1 Augmentation moyenne du C-index après pré-filtrage et niveau de significativité pour les 16 cancers étudiés. 101

4.1 Caractéristiques des onze cancers étudiés dans le chapitre 4. 115
4.2 Variables cliniques présentes pour chacun les 11 cancers étudiés. 121
4.3 Taille de banque médiane conseillée pour les 11 cancers pour les données RNA-seq de miARN. 128
4.4 Taille de banque médiane conseillée pour les 11 cancers pour les données mRNA-seq. 128

A.1 Différences médianes entre les p-valeurs oracles du modèle de Cox univarié (-log10) et estimés par le modèle de Cox pénalisé pour les 26 cancers de TCGA étudiés. VII
A.2 Différences médianes entre les IBS estimés par le modèle de Cox pénalisé et les IBS orocles pour les 16 cancers de TCGA étudiés. VIII
A.3 Différences médianes entre les p-valeurs (-log10) obtenues si la vérité terrain était connue et les p-valeurs (-log10) estimées pour les 26 cancers de TCGA. X

A.4 Différences médianes entre les IBS obtenus si la vérité terrain était connue et les IBS estimées pour les 26 cancers de TCGA. XI
A.5 Augmetation médiane de la p-valeur du modèle de Cox univarié (-log10) après pré-filtrage et niveau de significativité pour les 26 cancers étudiés. . . XII
A.6 Diminution médiane de l'IBS après pré-filtrage et niveau de significativité pour les 26 cancers étudiés. XII

Contributions

Conférences

— ***Post-selection inference and multiple testing*** - 7-9 février 2018 - Toulouse (France)
Centre International de Mathématiques et d'Informatique de Toulouse
Auditeur.

— **Journées annuelles 2019 du CLARA - 4-5 avril 2019** - Lyon (France)
Cancéropôle Lyon Auvergne Rhônes-Alpes
Présentation sous forme d'un poster.

— **ISMB/ECCB 2019** - Bâle (Suisse)
International conference on Intelligent Systems for Molecular Biology (ISMB) / European Conference on Computational Biology (ECCB)
Présentation sous forme d'un poster.

— **GRETSI** - 26-29 août 2019 - Lille (France)
Colloque francophone de traitement du signal et des images
Présentation sous forme de poster.

— ***Health Data Challenge (2nd edition) : Matrix factorization and deconvolution methods to quantify tumor heterogeneity in cancer research*** - 25-29 novembre 2019 - Aussois (France)
Data Institute - Université Grenoble Alpes
Présentation sous forme d'un poster.

Publications

— *Bioinformatics Methods to Select Prognostic Biomarker Genes from Large Scale Datasets : A Review.* Rémy Jardillier, Florent Chatelain, Laurent Guyon. Biotechnology Journal, Wiley-VCH Verlag, 2018, 13 (12), pp.1800103. ⟨10.1002/biot.20180010⟩

— *Benchmark of lasso-like penalties in the Cox model for TCGA datasets reveal improved performance with pre-filtering and wide differences between cancers.* Rémy Jardillier, Florent Chatelain, Laurent Guyon. bioRxiv, 2020. ⟨10.1101/2020.03.09.984070⟩

— (En cours d'écriture) *Cancer prognosis with miRNA-seq data and the Cox model : from few thousands to more than 5 millions reads needed.* Rémy Jardillier, Florent Chatelain, Laurent Guyon.

— (En cours d'écriture) *Issues in cancer prognosis with RNA-seq data and the Cox model : an overview from a simulation study.* Rémy Jardillier, Florent Chatelain, Laurent Guyon.

Chapitre 1

Introduction

Sommaire

1.1 Cancer, recherche, et prédiction de la survie . . . 3
1.1.1 Le cancer en chiffres . . . 3
1.1.2 Cancer et recherche . . . 3
1.1.3 Le « principe d'unicité du cancer » . . . 4
1.1.4 Médecine stratifiée et prédiction de la survie . . . 5
1.2 Les molécules d'ARN . . . 5
1.2.1 La notion de biomarqueurs . . . 5
1.2.2 Synthèse des ARN . . . 6
1.2.3 Les différents types d'ARN . . . 6
1.2.4 ARN et cancer . . . 7
1.3 Le séquençage : mesure du niveau d'expression des gènes . . . 8
1.3.1 La puce à ADN, première technique de mesure du niveau d'expression des gènes . . . 8
1.3.2 Principe général du séquençage haut débit . . . 8
1.3.3 Préparation des banques . . . 9
1.3.4 Amplification PCR . . . 9
1.3.5 Séquençage des banques . . . 10
1.3.6 Alignement sur un génome de référence . . . 10
1.3.7 Taille des banques et profondeur de séquençage . . . 10
1.4 Normalisation des données de comptage RNA-seq . . . 12
1.4.1 Normalisation CPM et pré-filtrage . . . 12
1.4.2 Normalisation log2-CPM . . . 13
1.4.3 Biais de la normalisation CPM et ajustement . . . 13
1.5 La base de données TCGA . . . 15
1.5.1 Description de la base de données . . . 15
1.5.2 Choix des cancers étudiés . . . 15
1.5.3 Correction de tests multiples . . . 16

1.6 Modélisation des données de survie 17
1.6.1 Les censures, singularité des données de survie 17
1.6.2 Notations pour les données de survie et les données génétiques. . . 17
1.6.3 Approche non-paramétrique : l'estimateur de Kaplan-Meier 18
1.6.4 Approche semi-paramétrique : le modèle de Cox 19
1.7 Métriques d'évaluation de la qualité de prédiction 21
1.7.1 L'indice pronostique 21
1.7.2 Procédure d'évaluation des modèles prédictifs 22
1.7.3 La concordance 23
1.7.4 La p-valeur du modèle de Cox univarié 25
1.7.5 Le score de Brier 25
1.7.6 Remarques sur les scores d'évaluation des performances de prédiction 26
1.8 La « malédiction de la grande dimension » 28
1.8.1 Définition 28
1.8.2 Pré-filtrage univarié des gènes 28
1.8.3 La régression pénalisée 29
1.9 Conclusions 33
1.9.1 Tableau récapitulatif des données de survie et mRNA-seq de TCGA 33
1.9.2 Résumé 34
1.9.3 Objectifs de la thèse 34
1.9.4 Contexte de la thèse 35
1.10 Références 35

1.1 Cancer, recherche, et prédiction de la survie

1.1.1 Le cancer en chiffres

L'Institut National du Cancer (INCa) estime à 382 000 le nombre de nouveaux cas de cancer en France métropolitaine en 2018 (204 600 chez l'homme et 177 400 chez la femme), et à 157 400 le nombre de décès (89 600 chez l'homme et 67 800 chez la femme). Ces chiffres placent le cancer comme première cause de décès prématurés (décès avant 65 ans) en France depuis 2004. Les cancers les plus répandus sont ceux du sein chez la femme et de la prostate chez l'homme. Le cancer le plus meurtrier est celui du poumon, avec 33 000 décès en 2018 (10 000 chez la femme et 23 000 chez l'homme).

Sur la période 2010-2018 en France, le nombre de nouveaux cas (taux d'incidence standardisée selon l'âge) tend à se stabiliser chez la femme (+0,7% par an) et à diminuer chez l'homme (-1,4% par an). En revanche, le nombre de décès (taux de mortalité standardisé selon l'âge) est en baisse à la fois chez la femme et chez l'homme (-0,7% chez la femme et -2% chez l'Homme entre 2010 et 2018).

Toujours selon l'INCa, le tabac constituerait le facteur de risque qui engendre le plus de cancer (19,8% des nouveaux cas), suivi par l'alcool, l'alimentation désiquilibrée et le surpoids (respectivement 8%, 5,4% et 5,4% des nouveaux cas). Ainsi, il est estimé que 41% des cancers pourraient être évités en changeant son mode de vie.

En 2018 dans le monde, le nombre de nouveau cas est estimé à 18,1 millions, et le nombre de décès à 9,6 millions [Bray et collab., 2018]. L'Organisation Mondiale de la Santé (OMS) prévoit que si les tendances actuelles se poursuivent, le monde enregistrera une augmentation de 60% des cas de cancers au cours des deux prochaines décennies [WHO, 2020].

1.1.2 Cancer et recherche

En 1971, le président des États-Unis Richard Nixon prononce un discours aux connotations martiales devant le congrès et les caméras de télévision en déclarant « la guerre contre le cancer ». La National Cancer Act est alors signé, et le budget de la recherche alloué au cancer est multiplié par 10 (1,5 milliards de dollars). Après la conquête spatiale et l'avènement du nucléaire, le niveau de confiance dans le progrès scientifique est à son paroxysme. Bien que les objectifs du président (vaincre le cancer à l'horizon 1981) ne seront pas atteints [LAG et collab., 2008], ce plan historique a permis d'insuffler une dynamique de recherche et de coordination.

En France, les deux premiers instituts de lutte contre le cancer sont créés dans les années 1920. L'Institut Curie se concentre sur la radiothérapie, et l'Institut du cancer de Villejuif (devenu l'Institut Gustave Roussy) se focalise sur l'intégration des soins, de la recherche, et de l'enseignement.

À l'échelle gouvernementale, des initiatives telles que les « plans cancer » ont été mise

en place. Cette dernière, initiée en 2003 par Jacques Chirac, a pour but d'établir des politiques publiques de lutte contre le cancer. Ainsi, des mesures telles que l'augmentation du prix du tabac, des campagnes de dépistage du cancer du sein, la préservation de la qualité de vie, ou encore le financement de la prévention et de la recherche. Selon L'INCa, 180 millions d'euros ont été attribués à la recherche en 2017 en France, dont 115,95 millions d'euros par les organismes institutionnels, 36,45 millions d'euros par la Ligue contre le cancer, et 28,4 millions d'euros par la Fondation ARC pour la recherche sur le cancer.

1.1.3 Le « principe d'unicité du cancer »

Cette thèse porte sur les méthodes mathématiques pour le pronostic. Ainsi, les notions de biologie sont importantes pour une bonne compréhension du sujet et des enjeux, mais ne font pas partie intégrante du cœur de l'étude. A ce titre, les explications biologiques essentielles seront explicitées le long du manuscrit, mais resteront superficielles. Pour plus de détails, nous référons le lecteur au livre de PEZZELLA et collab. [2019] qui s'intéresse de manière détaillée à la biologie du cancer.

L'INCa définit le cancer comme une « maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive. Ces cellules déréglées finissent par former une masse qu'on appelle tumeur maligne » (`https://www.e-cancer.fr/Dictionnaire/C/cancer`). Une tumeur peut aussi être bénigne (*e.g.* polypes, verrues) et n'est alors pas considérée comme un cancer : elle se développe lentement, ne récidive pas si elle est enlevée, et ne peut pas produire de métastases.

Le mot « cancer » tire son origine du mot latin homonyme qui signifie crabe. C'est Hippocrate (460-377 avant J-C) qui le premier compare la forme des tumeurs du sein à ce crustacé [PAPAVRAMIDOU et collab., 2010]. Ces tumeurs ont effectivement des excroissances qui ressemblant aux pattes d'un crabe.

Chaque être humain possède des caractéristiques génétiques, épigénétiques, transcriptomiques, protéomiques, métabolomiques et microbiotiques uniques. De plus, le « microenvironnement tumoral », défini comme l'environnement (cellules, vaisseaux sanguins, molécules) autour de la tumeur, joue un rôle primordial dans le développement et l'agressivité des cancers [JUNTTILA et DE SAUVAGE, 2013], et diffère d'un patient à un autre. Ensuite, les tumeurs d'un même organe possèdent des caractéristiques génétiques et phénotypiques différentes entre les patients (« hétérogénéité inter-tumorale ») [SUN et YU, 2015] et au sein d'une tumeur (« hétérogénéité intra-tumorale ») [BERGER et collab., 2018; GERLINGER et collab., 2012]. Enfin, le mode de vie joue un rôle important dans la survenue et le développement des cancers [LOOMANS-KROPP et UMAR, 2019]. L'ensemble de ces facteurs contribuent à l'unicité de chaque cancer, et appuie la nécessité d'une prise en charge individualisée des patients [OGINO et collab., 2012].

1.1.4 Médecine stratifiée et prédiction de la survie

Les deux principaux évènements étudiés dans les modèles de survie et que l'on cherche à prédire sont :

— le décès du patient.

Dans ce cas, le temps de survie est défini comme le temps entre le diagnostic et le décès, et l'on parle alors de survie globale. Cependant, le décès peut être dû à une cause extérieur au cancer (*e.g.* accident de la route) et ainsi biaiser l'analyse et les résultats.

— la survie sans progression, définie comme le temps entre le diagnostic et l'apparition d'un nouvel évènement associé au cancer.

Dans notre étude, un nouvel évènement correspond à la progression de la maladie, une récidive loco-régionale, l'apparition de métastases, l'apparition d'une nouvelle tumeur primaire ou le décès du patient avec présence de la tumeur [Liu et collab., 2018a]. L'évènement étudié est donc directement associé au cancer, et les temps observés sont classiquement plus courts que pour la survie globale.

Dans un rapport publié en février 2014, la Haute Autorité de Santé définit la médecine stratifiée (ou médecine de précision) comme « une approche thérapeutique où l'objectif est de sélectionner les patients auxquels administrer un traitement en fonction d'un marqueur prédictif, afin de ne traiter que la sous-population susceptible de recevoir un bénéfice du traitement ».

Hood et Friend [2011] place la prévention, la stratification, la participation et la prédiction (« P4 ») comme piliers d'un nouveau modèle de prise en charge des patients atteints de cancer. En particulier, la prédiction de la survie globale ou de la récidive apparaît essentielle pour un meilleur suivi et une meilleure prise en charge des patients [Hagerty et collab., 2005; Rabin et collab., 2013]. Les molécules d'ARN jouent un rôle central dans le développement et l'agressivité des cancers, et nous nous attacherons à décrire leur rôle dans la prochaine partie.

1.2 Les molécules d'ARN

1.2.1 La notion de biomarqueurs

L'Institut National de la Santé et de la Recherche Médicale (INSERM) défini un biomarqueur comme « une molécule (enzyme, hormone, métabolite, etc.), voire un type de cellule, dont la présence ou la concentration anormale dans le sang ou les urines signale un évènement ou un statut physiologique particulier. »

Les biomarqueurs que nous utilisons sont les molécules d'ARN, et nous nous attacherons à décrire leur rôle dans l'organisme et la manière dont ils sont synthétisés dans les

parties suivantes. Pour plus de détails concernant la biologie de la cellule, nous référons le lecteur au livre ALBERTS et collab. [2018].

1.2.2 Synthèse des ARN

Ce travail porte sur des cancers humains; nous décrirons donc ce mécanisme dans les cellules eucaryotes (qui possèdent un noyau). L'acide désoxyribonucléique (ADN) est la principale molécule qui compose les chromosomes. Elle est formée de deux brins qui forment une hélice - on dit que l'ADN est « double brin ». Les brins sont reliés entre eux par l'intermédiaire des bases azotées (adénine (A), cytosine (C), guanine (G), thymine (T)), qui se lient entre elles par des liaisons hydrogènes (A avec T et C avec G). Un gène se définit alors comme une séquence d'ADN déterminée par sa position sur un chromosome.

La transcription est à l'origine de l'expression du génome d'un individu. Elle correspond à la copie simple brin d'un segment particulier de l'ADN, un gène, en acide ribonucléique (ARN) dans le noyau de la cellule (Fig. 1.1). L'ARN peut ainsi être vu comme le support intermédiaire de l'information contenu dans les gènes. On dit que les gènes « s'expriment », et l'on parle alors de « niveau d'expression du gène » pour faire référence à la quantité d'ARN transcrite.

L'ARN peut migrer dans le cytoplasme des cellules et nous nous attacherons à décrire leurs caractéristiques et leur rôle dans la prochaine section.

1.2.3 Les différents types d'ARN

Les molécules d'ARN ainsi transcrites, le transcriptome, peuvent avoir différentes fonctions, dont :

— permettre la synthèse des protéines (on parle alors d'ARN messager) (Fig. 1.1).

— intervenir dans la régulation des gènes.

— accomplir des fonctions catalytiques.

Dans cette étude, nous nous intéresserons à deux classes d'ARN particuliers, les ARN messagers (ARNm), et les microARN (miARN).

Les ARNm sont traduits en protéines dans le cytoplasme de la cellule. Ces protéines sont essentielles au bon fonctionnement des cellules et des tissus, et sont impliquées entre autres dans les réactions chimiques et de dégradations indispensables au métabolisme (*e.g.* enzymes), dans la structure des tissus (*e.g.* la kératine constitue l'essentiel de nos cheveux et de nos ongles), ou encore dans la régulation (*e.g.* l'insuline permet de réguler la glycémie). La taille des ARN messagers est très variable, pouvant aller de quelques dizaines à plusieurs milliers de nucléotides.

Les miARN sont de petites molécules d'ARN (20 à 24 nucléotides en général) non codantes (*i.e.* qui ne sont pas traduites en protéines) et qui interviennent dans la régulation

post-transcriptionnelle des gènes (*i.e.* les miARN agissent directement sur les molécules d'ARN) [BARTEL, 2018].

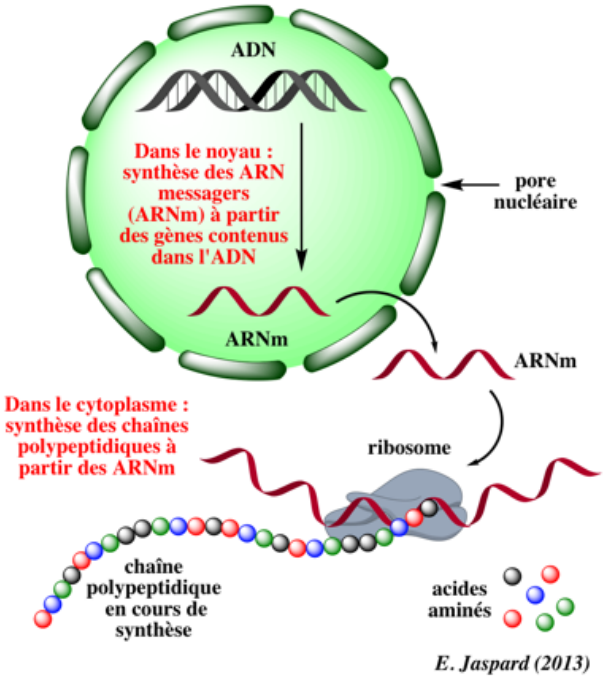


FIGURE 1.1 – **Transcription de l'ADN en ARN dans le noyau, et traduction de l'ARN en protéine dans le cytoplasme.**
Source : cours sur la synthèse des protéines d'Emmanuel Jaspard à l'université d'Angers (http://biochimej.univ-angers.fr/Page2/COURS/7RelStructFonction/2Biochimie/1SyntheseProteines/1Synthese Prot.htm).

1.2.4 ARN et cancer

Nous avons vu que les gènes sont à l'origine de nombreux mécanismes chimiques et biologiques au sein des cellules et des tissus. Ainsi, l'expression aberrante (*i.e.* expression anormalement élevée ou anormalement basse) de certains gènes peut-être à l'origine de certaines pathologies, notamment le cancer. Ce lien entre ARN et cancer a été démontré à la fois pour les ARN messagers [SORLIE et collab., 2001] et les miARN [CALIN et CROCE, 2006], et l'étude du niveau d'expression des gènes est intéressante pour mieux comprendre la biologie du cancer et pour des applications cliniques à la fois pour les ARNm [DUMBRAVA et MERIC-BERNSTAM, 2018] et pour les miARN [RUPAIMOOLE et SLACK, 2017].

Les gènes dont l'expression favorise la survenue et/ou le développement des tumeurs sont dits « oncogènes ». A l'inverse, les gènes dont l'expression prévient la survenue et/ou le développement des tumeurs sont dits « tumeurs suppresseurs ».

L'étude du transcriptome d'un morceau de tumeur est rendu possible par le séquençage. Nous nous attacherons à décrire cette technologie dans la partie suivante.

1.3 Le séquençage : mesure du niveau d'expression des gènes

1.3.1 La puce à ADN, première technique de mesure du niveau d'expression des gènes

La première technique qui a permis de mesurer le niveau d'expression des gènes à grande échelle est la « puce à ADN » (*microarray* en anglais) [FODOR et collab., 1991; SCHENA et collab., 1995]. Le principe de cette technologie repose sur différentes étapes :

1. isoler l'ARNm de l'échantillon à étudier et le transformer en ADN simple brin.
2. ajouter un élément fluorescent à chaque brin.
3. placer l'ensemble des brins sur une puce qui contient des puits avec des ADN simples brin composés des nucléotides des gènes à mesurer.
4. laver pour éliminer l'ensemble des brins issus de l'échantillon qui ne sont pas associés aux brins de la puce.
5. mesurer l'intensité de fluorescence de chaque puits afin de quantifier le niveau d'expression des gènes.

Cette technologie a permis la révolution de la génomique, notamment dans le domaine de la prédiction de la survie de patients atteints de cancer [BEER et collab., 2002; GUI et LI, 2005; VAN DE VIJVER et collab., 2002]. En revanche, ces puces à ADN ne permettent pas d'avoir un aperçu de l'ensemble du transcriptome d'un échantillon, mais seulement d'un ensemble de gènes pré-définis par l'utilisateur [RAO et collab., 2019]. La technologie « RNA-Seq » (*RNA-sequencing*) permet de remédier à cela et autorise la découverte de nouveaux transcrits [WANG et collab., 2009]. Nous nous attacherons à la décrire dans les paragraphes ci-dessous.

1.3.2 Principe général du séquençage haut débit

La technique du séquençage haut débit (HTS - *High Throughput Sequencing*) [REUTER et collab., 2015] permet de quantifier la quantité de molécules d'ARN transcrites par le génome dans un échantillon à un instant donné. Plus précisément, la technologie RNA-Seq est utilisé pour produire la base de données TCGA (paragraphe 1.5.1) que nous utilisons. On parle de « mRNA-seq » lorsque cette technologie est appliquée aux ARN messagers, et de « miRNA-seq » lorsqu'elle est appliquée aux miARN (`https://docs.gdc.cancer.gov/Data/Introduction/`, `http://cancergenome.nih.gov/cancersselected/biospeccriteria`, CHU et collab. [2016]).

Dans cette partie, nous allons présenter le protocole de l'entreprise Illumina (`https://www.illumina.com/`, MEYER et KIRCHER [2010]), très utilisé. Cette méthode comporte quatre étapes distinctes (Fig. 1.2) :

1. la préparation d'une « banque » pour chaque échantillon (biopsie de la tumeur dans notre cas).
2. l'amplification PCR (acronyme signifiant « *Polymerase Chain Reaction* »).
3. le séquençage des banques.
4. l'alignement sur un génome de référence.

Nous nous attacherons à décrire ces quatre étapes dans les deux paragraphes suivant.

1.3.3 Préparation des banques

Une banque est une collection de fragments d'ADN. Le but de cette première étape est de créer une banque pour chaque échantillon afin de procéder à la quantification grâce au séquençage.

Cette étape se décompose en quatre tâches (Fig. 1.2.A) :

1. isoler l'ARN présent dans l'échantillon. La biopsie de la tumeur est un ensemble de cellules et de liquides organiques qui doivent être dégradés puis purifiés pour ne garder que l'ARN.
2. fragmenter les ARN en morceaux. Les machines d'Illumina permettant le séquençage peuvent séquencer des molécules ne contenant qu'au plus 200 / 300 bases, alors que les molécules d'ARN peuvent être composées de plusieurs milliers de bases.
3. convertir les fragments d'ARN en ADN simple brin. Ces molécules sont plus stables que les molécules d'ARN, et donc plus robustes et faciles à manipuler.
4. ajouter des adaptateurs à chaque fragments d'ADN. Ces adaptateurs sont de petites séquences d'ADN qui vont venir prolonger les fragments d'ADN obtenus à l'étape précédente. Chaque adaptateur est spécifique d'un échantillon donné ; ils jouent le rôle de carte d'identité de ces fragments, et permettent donc de séquencer différents échantillons en même temps.

Cette étape a donc permis de passer d'une biopsie de tumeur à un échantillon comportant des fragments d'ADN jumelés à des cartes d'identité, les adaptateurs. Cet échantillon est appelé la « banque ». L'objectif des deux prochaines étapes est de quantifier la quantité d'ARN présente en séquençant l'ensemble des banques.

1.3.4 Amplification PCR

Les fragments d'ADN sont placés sur une grille (*flow cell*) avant d'être amplifiés (Fig. 1.2.B). L'amplification PCR permet de copier chaque fragment d'ADN localement afin de créer des *clusters*, et donc d'augmenter la sensibilité du signal reçu lors du séquençage.

Chaque grille peut contenir des fragments d'ADN de différents échantillons, et cette caractéristique est appelée le « multiplexage ». Suivant le nombre de *clusters* que peut contenir la grille et le nombre de patients présents dans la cohorte, différentes grilles vont être utilisées.

1.3.5 Séquençage des banques

Le séquençage consiste à déterminer la séquence de nucléotides de chaque fragment (Fig. 1.2.C). Pour cela, des sondes fluorescentes sont attachés au premier nucléotides de chaque fragment. Ces sondes sont spécifiques : une couleur différente pour chacun des nucléotides A, C, G, T. Une image de fluorescence est enregistrée, et cette opération est répétée séquentiellement, pour chaque nucléotide jusqu'à la fin des fragments. Ainsi, la séquence de chaque fragment présent dans chaque banque est connue. Les séquences ainsi générées de chacun de ces fragments sont appelées « lectures ».

Par abus de langage, le terme de séquençage fait référence à l'ensemble du processus permettant de mesurer le niveau d'expression des gènes (RNA-seq).

1.3.6 Alignement sur un génome de référence

Pour quantifier l'expression des gènes dans un échantillon, les lectures sont alignées sur un génome de référence contenant les séquences de nucléotides de l'ensemble des gènes (Fig. 1.2.D). Le niveau d'expression d'un gène se traduit alors par le nombre de lectures qui se sont alignées sur une partie de sa séquence. On parle de « données de comptage ».

1.3.7 Taille des banques et profondeur de séquençage

Pour un échantillon donné, la « taille de la banque » peut se définir de deux manières :

— le nombre total de lectures séquencées.

— le nombre total de lectures séquencées qui ont été alignés sur le génome de référence.

Dans la suite, nous utiliserons la deuxième définition.

Il est important de noter que la taille des banques peut varier d'un patient à l'autre pour un même protocole. Par exemple, quelques fragments d'ADN ne reçoivent pas d'adaptateurs, et ne vont pas être séquencés. De plus, certaines lectures ne sont pas alignées sur le génome (*e.g.* problème de séquençage de l'adaptateur qui empêche la lecture d'être associée à un patient). Ensuite, l'amplification PCR de séquences d'ADN riches en bases nucléotidiques guanine (G) et cytosine (C) est difficile [MAMMEDOV et collab., 2008], et implique un biais dans les données RNA-seq [BENJAMINI et SPEED, 2012; RISSO et collab., 2011]. Enfin, le nombre de fragments peut varier d'une grille à l'autre, et le multiplexage

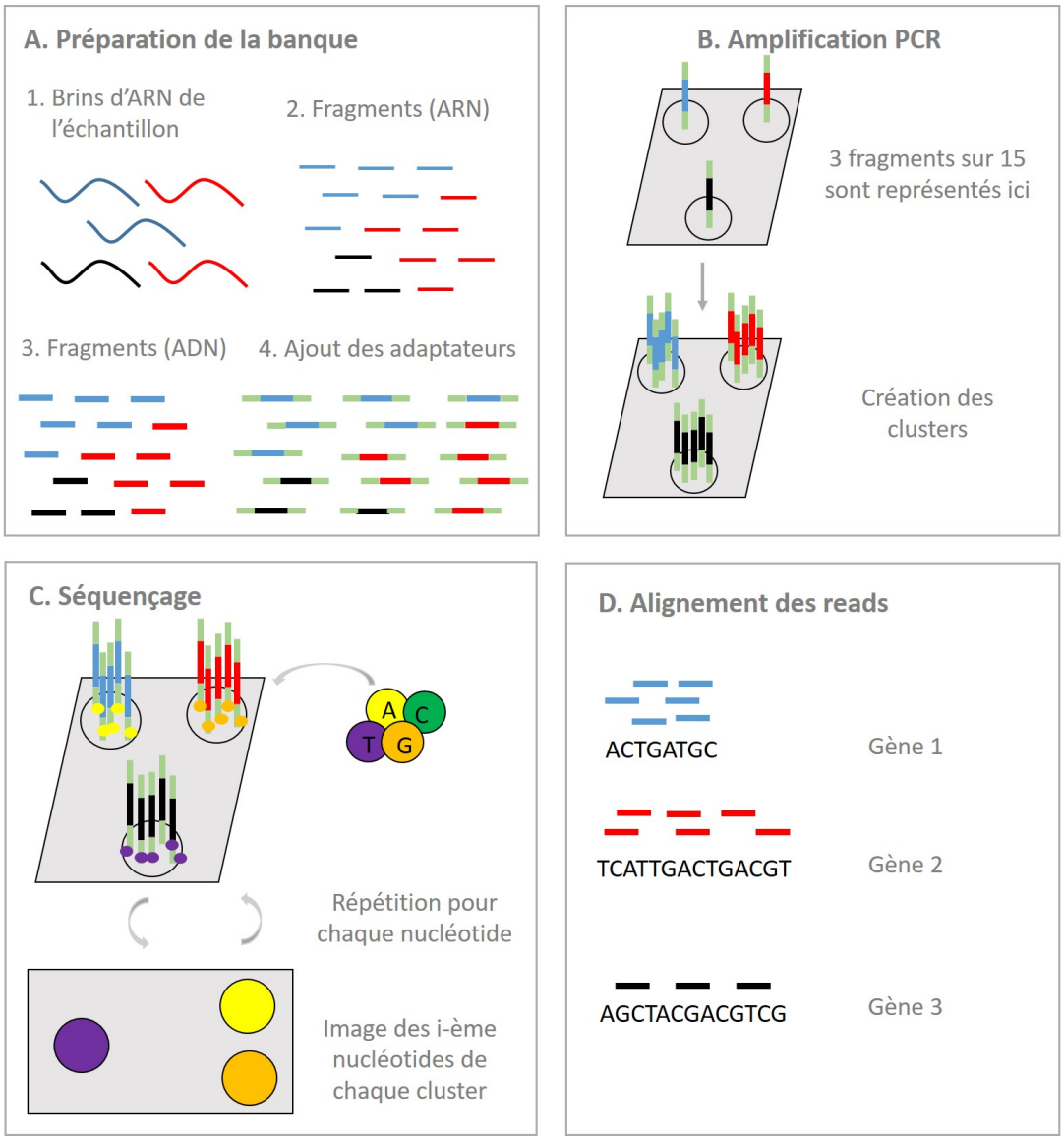


FIGURE 1.2 – **Différentes étapes du séquençage haut débit (RNA-seq)**
(A) Chaque couleur (bleu, rouge, noir) correspond à des séquences d'ARN différentes. Pour chaque échantillon (biopsie d'une tumeur), les molécules d'ARN sont découpées en fragments, transformées en simple brin, et des adaptateurs sont ajoutés. Ces derniers permettent de les identifier, et les différents échantillons peuvent ainsi être mélangés pour être séquencés en parallèle. (B) Les fragments sont placés sur une grille (*flow cell*) et amplifiés (PCR) pour former des *clusters*. Le signal obtenu lors de l'étape suivante sera ainsi plus important. (C) Des sondes fluorescentes spécifiques d'un nucléotide (A - jaune, C - vert, G - orange, ou T - violet) sont ajoutées et vont venir s'attacher au premier nucléotide de chaque fragment. Une image est alors prise, et cette procédure est répétée jusqu'à ce que les séquences nucléotidiques de chaque fragment soient déterminées. (D) Les séquences ainsi obtenues, les lectures, sont alors alignées sur un génome de référence pour déterminer de quel gène ils sont issus.

peut induire des nombres de lectures différentes pour deux patients dont les échantillons ne sont pas sur la même grille.

La « profondeur de séquençage » (ou couverture) moyenne (c) se calcule à partir de la

longueur du génome de référence (G), de taille de la banque (N), et de la taille moyenne des lectures (L). Elle se définit comme [LANDER et WATERMAN, 1988] :

$$c = \mathrm{N} \times \frac{\mathrm{L}}{\mathrm{G}}$$

Ainsi, la profondeur de séquençage est une fonction croissante de la taille de la banque. Pour un ensemble d'échantillons qui ont été séquencés suivant le même protocole, le facteur $\frac{\mathrm{L}}{\mathrm{G}}$ est le même, et les notions de « profondeur de séquençage » et de « taille des banques » peuvent être utilisés de manière équivalente.

La profondeur de séquençage est un paramètre clef de l'analyse du génome : plus elle est importante, plus les mesures seront précises et fines : le nombre de gènes détectés sera plus important, mais l'analyse sera plus coûteuse. Bien que les coûts de la technologie RNA-seq aient diminué au cours de la dernière décennie, ils restent importants et la technologie RNA-seq est peu utilisée pour le moment en clinique [CIEŚLIK et CHINNAIYAN, 2018; KUMAR-SINHA et CHINNAIYAN, 2018; SENFT et collab., 2017]. Il apparaît donc essentiel de déterminer une profondeur de séquençage suffisante pour obtenir des résultats reproductibles et optimaux, mais pas trop élevée pour ne pas faire exploser les coûts [SIMS et collab., 2014]. Il s'agit donc d'optimiser les prédictions et le nombre d'échantillons séquencés sous contrainte de coûts.

Ainsi, MILANEZ-ALMEIDA et collab. [2020] ont montré qu'il était possible de réduire la taille des banques de la base de données TCGA d'un facteur 100 pour les ARN messagers sans dégrader les qualités de prédiction. L'étude de l'impact de la profondeur de séquençage des miARN sur la prédiction fera l'objet du chapitre 4.

1.4 Normalisation des données de comptage RNA-seq

1.4.1 Normalisation CPM et pré-filtrage

La taille des banques est différente suivant les patients, et cela introduit un biais dans la comparaison du niveau d'expression d'un gène entre les patients. En effet, séquencer un échantillon avec une profondeur deux fois moins importante qu'un autre va conduire, en moyenne, à un nombre deux fois moins important de lectures alignées sur chaque gène. La normalisation *Count Per Million* (CPM) permet remédier à ce problème en divisant les données de comptage d'un patient i pour un gène j par la taille de la banque :

$$\mathrm{CPM}_{ij} = \frac{\mathrm{R}_{ij}}{\mathrm{R}_i} \times 10^6,$$

avec CPM_{ij} la donnée CPM-normalisée, R_{ij} la donnée de comptage, et R_i la taille de la banque du patient i.

Les gènes retenus sont ceux pour lesquels les niveaux d'expression CPM-normalisés sont supérieurs à 1 pour au moins 1% des patients. La fonction `cpm` du package `edgeR`

[ROBINSON et collab., 2010] permet d'effectuer cette normalisation CPM.

1.4.2 Normalisation log2-CPM

La normalisation log2-CPM se définit de la manière suivante :

$$X_{ij} = \log 2\left(\frac{R_{ij} + 0.5}{R_i + 1} \times 10^6\right),$$

avec X_{ij} la donnée normalisée (log2-CPM), R_{ij} la donnée de comptage, et R_i la profondeur de séquençage du patient i.

Deux différences sont à noter par rapport à la formule de la normalisation CPM présentée dans le paragraphe précédent :

— le nombre de lectures R_{ij} est incrémenté de 0,5 pour éviter de prendre le logarithme de 0 et réduire la variabilité de la normalisation log2-CPM pour les gènes faiblement exprimés.

— la taille de la banque est incrémentée de 1 pour s'assurer que la quantité $\frac{R_{ij}+0.5}{R_i+1}$ soit comprise entre 0 et 1.

1.4.3 Biais de la normalisation CPM et ajustement

Dans le cas où plusieurs gènes seraient surexprimés (resp. sous-exprimés) entre deux patients, la normalisation CPM introduit un biais. Prenons l'exemple de deux patients A et B pour lesquels trois gènes sont séquencés avec la même profondeur (Fig. 1.3). Imaginons que les trois gènes s'expriment de manière équivalente chez le patient A (*i.e.* même quantité d'ARN transcrite par chacun des gènes), et que les gènes du patients B ont le même niveau d'expression que ceux du patient A, sauf pour le troisième gène qui ne s'exprime pas. Dans ce cas, les données CPM-normalisées seront de $3,3 \times 10^5$ pour les trois gènes du patient A, de $5,0 \times 10^5$ pour les deux premiers gènes du patient B, et de 0 pour le dernier. Ainsi, la présence d'un gène différentiellement exprimé biaise les résultats de la normalisation CPM : les deux premiers gènes ont des niveaux d'expression CPM-normalisés différents, alors que nous avons supposé que leur niveau d'expression réel était les mêmes chez les deux patients.

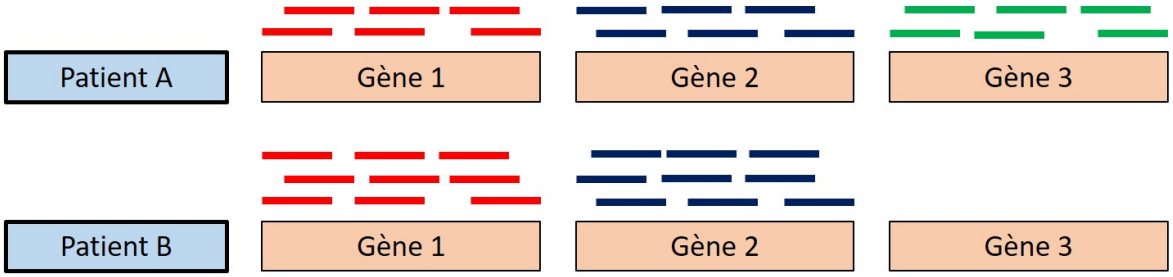
La méthode TMM (*Trimmed Mean of M values*, ROBINSON et OSHLACK [2010]) permet de remédier à ce biais. Elle se décompose en quatre étapes :

— choix d'un patient de référence i'.

— calcul d'un facteur de normalisation M_{ij} pour chaque gène et chaque patient.

$$M_{ij} = \frac{CPM_{ij}}{CPM_{i'j}}$$

A



B

	Gène 1	Gène 2	Gène 3
Patient A	3,3 . 10^5	3,3 . 10^5	3,3 . 10^5
Patient B	5, 0 .10^5	5,0 . 10^5	0

FIGURE 1.3 – **Exemple de biais induit par la normalisation CPM**
(A) Alignement des lectures sur un génome de référence pour deux patients et trois gènes. Les trois gènes s'expriment de manière équivalente chez le patient A (*i.e.* même quantité d'ARN transcrite par chacun des gènes), et les gènes du patients B ont le même niveau d'expression que ceux du patient A, sauf pour le troisième gène qui ne s'exprime pas. La taille des banques est la même pour les deux patients (18 lectures alignées sur le génome).
(B) Niveaux d'expression CPM-normalisés pour les trois gènes chez les deux patients. Malgré des niveaux d'expression réels similaires pour les deux premiers gènes chez les deux patients, les niveaux d'expression CPM-normalisés sont différents. La présence du troisième gène différentiellement exprimé induit un biais.

— pour chaque patient i, un facteur de normalisation moyen M_i est calculé en ne gardant que les coefficients M_{ij} compris entre le quantile 15 et le quantile 85. Cela permet d'exclure les gènes différentiellement exprimés entre le patient de référence et le patient i.

— ce facteur de normalisation moyen est alors utilisé pour normaliser la taille de la banque du patient i : $\tilde{R}_i = R_i \times M_i$. On obtient alors :

$$\tilde{X}_{ij} = \log 2\left(\frac{R_{ij} + 0.5}{\tilde{R}_i + 1} \times 10^6\right),$$

avec $\tilde{R}_i$ la taille de la banque normalisée.

La fonction `calcNormFactors` du package `edgeR` [ROBINSON et collab., 2010] permet de calculer le facteur de normalisation pour chaque patient, et la fonction `voom` du package `limma` [RITCHIE et collab., 2015] permet d'effectuer cette normalisation CPM ajustée par la méthode TMM.

1.5 La base de données TCGA

1.5.1 Description de la base de données

The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) est une base de données publiques multicentriques qui provient de la collaboration entre deux instituts américains, le *National Cancer Institute* (NCI) et le *National Human Genome Research Institute* (NHGRI). Ce projet a été lancé en 2005 et a pour but d'obtenir et d'analyser le transcriptome et le génome de tissus tumoraux et de quelques tissus sains environnant la tumeur obtenus à partir de biopsies. A ce jour, plus de 20 000 échantillons provenant de plus de 11 000 patients et de 33 types de cancer ont été collectés, caractérisés et analysés.

Les données génomiques et transcriptomiques sont associées à des données cliniques et de survie (*e.g.* âge, sexe, taille de la tumeur, présence de métastases ou non, temps de suivi du patient après le diagnostic, décès observé au cours du suivi ou non). Différents évènements et temps de survie associée peuvent être étudiés à partir des données cliniques. Liu et collab. [2018b] ont analysé les données de survie TCGA et donnent des recommandations sur l'utilisation de différents évènements, dont la survie globale et la survie sans progression. Pour cela, une procédure permettant d'évaluer la qualité des données de survie a été mise en place, et des recommandations sont faites pour chaque cancer. L'analyse du transcriptome des tumeurs et des données de survie a permis l'émergence de nouveaux biomarqueurs et de nouvelles thérapies pour des utilisations cliniques [Dumbrava et Meric-Bernstam, 2018].

1.5.2 Choix des cancers étudiés

Les analyses que nous allons présenter dans les chapitres suivants sont menées sur différents cancers. Nous avons décrit les noms des cancers et quelques caractéristiques des données dans le tableau 1.2. Dans la suite du manuscrit, nous utiliserons les acronymes de TCGA pour nommer les cancers, et nous nous référerons à ce tableau pour leur signification.

Le premier critère retenu pour le choix des cancers est le nombre de patients. Nous avons retiré les cancers pour lesquels les ARNm (resp. miRNA) ont été séquencés pour moins de 80 patients dans les études menées par TCGA. Ainsi, les cancers dont les acronymes sont DLBC, KICH, CHOL et UCS ne seront pas étudiés pour les ARNm (resp. DLBC, KICH, CHOL, GBM et UCS pour les miARN). Ensuite, Liu et collab. [2018b] recommandent de ne pas utiliser le mélanome cutané (SKCM - *Skin Cutaneous Melanoma*) car les données transcriptomiques de tumeurs primaires et de métastases sont fournies sans distinction dans TCGA. De plus, les auteurs suggèrent que les phéochromocytomes et les paragangliomes (PCPG - *Pheochromocytoma and Paraganglioma*) ne possèdent pas assez d'évènements et nécessitent des temps de suivi plus importants à la fois pour la survie globale et la survie sans progression. Toujours selon cette même étude, parmi les cancers

restants, la survie globale est utilisée, sauf pour ceux dont l'acronyme est BRCA, LGG, PRAD, READ, TGCT, THCA et THYM pour lesquels la survie sans progression est préférée [Liu et collab., 2018b]. Après cette première étape de choix, 26 cancers sont retenus pour les ARNm et 25 pour les miARN.

Tableau 1.1 – **Cancers retenus et données de survie utilisées.**
Le cancer GBM n'est pas utilisé pour les données miRNA-seq, et est indiqué en gras dans le tableau. Les données de séquençage des miARN ne sont pas disponible pour ce cancer.
« OS » : *Overall Survival* (*i.e* la survie globale est utilisée comme temps de survie).
« PFI » : *Progression Free Interval* (*i.e* la survie sans récidive est utilisée comme temps de survie).

Cancer	ACC	BLCA	BRCA	CESC	COAD	ESCA	**GBM**	HNSC	KIRC	KIRP	LAML	LGG	LIHC
Survie	OS	OS	PFI	OS	OS	OS	OS	OS	OS	OS	OS	PFI	OS
Cancer	LUAD	LUSC	MESO	OV	PAAD	PRAD	READ	STAD	TGCT	THCA	THYM	UCEC	UVM
Survie	OS	OS	OS	OS	OS	PFI	PFI	OS	PFI	PFI	PFI	OS	OS

1.5.3 Correction de tests multiples

Dans la suite, nous allons appliquer différents algorithmes à ces 26 cancers et, suivant l'objectif du chapitre, des p-valeurs seront calculées pour tester le niveau de significativité des résultats obtenus. Ainsi, dans ce contexte de tests multiples (une p-valeur sera assignée à chaque cancer), la probabilité de rejeter H_0 à tort au moins une fois tend vers 1 exponentiellement lorsque le nombre de tests tend vers l'infini :

$$\begin{aligned} P_{H_0^1,\dots,H_0^p}(\exists i \in 1,\dots,p, H_0^i \text{ est rejetée}) &= 1 - P_{H_0^1,\dots,H_0^p}(\forall i \in 1,\dots,p, H_0^i \text{ est acceptée}) \\ &= 1 - \prod_{i=1}^{p} P_{H_0^i}(H_0^i \text{ is accepted}) \\ &= 1 - (1-\alpha)^p \xrightarrow[p\to\infty]{} 1 \end{aligned}$$

Nous voyons ainsi qu'au risque α de 5% et pour les 26 cancers testés, nous avons un peu plus de 70% de chance de rejeter H_0 à tort pour au moins l'un d'eux. Pour remédier à cet écueil, Benjamini et Hochberg [1995] suggèrent de corriger les p-valeurs afin de contrôler le taux de rejets à tort (*i.e.* taux de fausses découvertes). Pour un certain taux α donné, cette méthode consiste à rejeter $H_0^{(1)},\dots,H_0^{(k)}$ pour le plus grand k tel que $p_{(k)} \leq \alpha\frac{k}{p}$, avec $p_{(1)},\dots,p_{(k)},\dots,p_{(p)}$ les p-valeurs classées par ordre croissant, et $H_0^{(1)},\dots,H_0^{(k)},\dots,H_0^{(p)}$ les hypothèses nulles associées.

Ainsi, lorsque nous étudierons les 26 cancers simultanément, une correction de tests multiples par la méthode de Benjamini-Hochberg présentée ci-dessus sera faite systématiquement, et nous le préciseront à chaque fois. Dans la partie suivante, nous nous attacherons à décrire les modèles mathématiques classiques permettant d'analyser ces données de survie.

1.6 Modélisation des données de survie

Dans cette partie, nous nous attachons à décrire les principaux outils mathématiques permettant de modéliser les données de survie. Pour plus de détails, nous référons le lecteur au livre « *The Statistical Analysis of Failure Time Data* » [KALBFLEISCH et PRENTICE, 2011].

1.6.1 Les censures, singularité des données de survie

On parle de « censure » (ou donnée censurée) lorsque l'évènement étudié (*i.e.* décès du patient ou nouvel évènement associé à la tumeur) n'a pas lieu au cours du suivi. Une censure est typiquement observée lorsque le patient guéri, quitte l'étude, ou lorsqu'il change d'hôpital et sort de la cohorte étudiée. Le « temps de censure » est alors défini comme le temps entre le diagnostic et la censure. Les données de survie sont donc bidimensionnelles et contiennent une variable de temps (le temps de suivi) et une variable binaire indiquant si l'évènement est observé ou non (par convention, 1 si l'évènement est observé, 0 sinon).

Dans le cas d'une censure, le temps entre le diagnostic et l'évènement n'est pas connu, mais le fait que l'évènement n'a pas été observé durant le suivi du patient est une information importante qui doit être prise en compte dans les modèles. L'enjeu des modèles de survie réside donc dans la prise en compte de toute l'information présente dans ces données hétérogènes censurées et non censurées. L'estimateur de Kaplan-Meier et le modèle de Cox sont deux approches largement utilisées pour étudier des données de survie et prendre en compte les censures, et nous nous attacherons à les décrire dans les paragraphes suivant.

1.6.2 Notations pour les données de survie et les données génétiques.

Dans la suite du texte, on notera T la variable aléatoire qui décrit le temps de survie, C la variable aléatoire associée au temps de censure, et δ le statut :

— $\delta = 1$ si $T \leq C$ (*i.e.* l'évènement est observé, donnée non-censurée).

— $\delta = 0$ si $T > C$ (*i.e.* la donnée est censurée).

Ainsi, une donnée de survie est définie comme le couple $(\min(T,C), \delta)$, où $\min(T,C)$ correspond au temps de suivi mesuré à partir du diagnostic. Enfin, le vecteur aléatoire de taille p modélisant les données génétiques (p gènes) sera noté $\mathbf{X} = (X_1, \ldots, X_p)^T$.

Lorsque l'on étudie un jeu de données, chaque patient correspond à une observation des variables aléatoires décrites précédemment. Ainsi, pour un jeu de données contenant n patients et p gènes,

— les données de survie correspondent à un vecteur $((t^{(1)}, \delta^{(1)}), \ldots, (t^{(n)}, \delta^{(n)}))^T$, avec $t^{(i)}$, le temps de suivi du patient i, et $\delta^{(i)}$ le statut du patient i ;

— les données génétiques correspondent à une matrice $(X_{ij})_{1\leq i\leq n, 1\leq j\leq p}$, où $\mathbf{X}^{(i)} = (X_{i1}, \ldots, X_{ip})$ est un vecteur qui correspond au niveau d'expression des p gènes pour le patient i, $i = 1, \ldots, n$.

1.6.3 Approche non-paramétrique : l'estimateur de Kaplan-Meier

La fonction de survie S à un temps t fixé est définie comme la probabilité de survivre jusqu'à l'instant t, c'est-à-dire :

$$S(t) = P(T > t), t \geq 0$$

L'estimateur de Kaplan-Meier [KAPLAN et MEIER, 1958] permet d'estimer cette fonction de survie. L'intuition qui permet de définir cet estimateur est la suivante : pour trois temps t, t' et t'' tel que $t > t' > t''$,

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t', T > t) \\ &= P(T > t | T > t') \times P(T > t') \\ &= P(T > t | T > t') \times P(T > t' | T > t'') \times P(T > t'') \end{aligned}$$

En remplaçant t' et t'' par l'ensemble des temps de suivi t_i de la cohorte de patients étudiés tel que $t_i < t$, on obtient :

$$S(t) = \prod_{i=1}^{k} P(T > t_{(i)} | T > t_{(i-1)}),$$

avec k le nombre de temps de suivi plus petits que t, $t_{(1)}, \ldots, t_{(k)}$ les temps de suivi inférieurs à t tel que $t_{(1)} \leq \ldots \leq t_{(k)}$, et $t_0 = 0$.

En notant $n_{(i)}$ le nombre de patients à risque au temps $t_{(i)}$ (*i.e.* ensemble des patients qui ont un temps de suivi strictement supérieur à $t_{(i)}$), et $d_{(i)}$ le nombre de décès en $t_{(i)}$, la probabilité $p_i = P(T \leq t_{(i)} | T > t_{(i-1)})$ d'observer l'évènement dans l'intervalle $]t_{(i-1)}, t_{(i)}]$ sachant qu'il n'avait toujours pas été observé en $t_{(i-1)}$ peut être estimée par $\frac{d_{(i)}}{n_{(i)}}$. On obtient ainsi l'estimateur de Kaplan-Meier de la fonction de survie :

$$\begin{aligned} \hat{S}(t) &= \prod_{t_i < t} (1 - p_i) \\ &= \prod_{t_i < t} \frac{n_i - d_i}{n_i} \end{aligned}$$

La fonction `survfit` du package `survival` [THERNEAU, 2020] permet de calculer l'estimateur de Kaplan-Meier. Cet estimateur $\hat{S}(t)$ est une fonction en escalier décroissante

du temps, les paliers arrivant aux instants ou l'évènement étudié est observé. On peut noter que l'estimateur de Kaplan-Meier est très général car il ne repose pas sur un modèle (paramétrique) particulier de distribution des temps de suivi ou de survie, et il est ainsi qualifié de non-paramétrique. Il est classiquement utilisé pour :

— avoir un aperçu de l'agressivité de la maladie étudiée (*e.g.* calcul du temps de survie médian, probabilité de survie à 3 ans), comme illustré Figure 1.4.

— comparer la survie de patients ayant reçus deux traitements différents. Le test du *log-rank* [MANTEL, 1966] permet alors de déterminer si les courbes de survie estimées par la méthode de Kaplan-Meier sont significativement distinctes.

L'estimateur de Kaplan-Meier permet de prendre en compte les censures, mais ne permet pas de relier analytiquement des variables explicatives aux données de survie. En ce sens, on peut qualifier cette méthode de descriptive et non d'explicative.

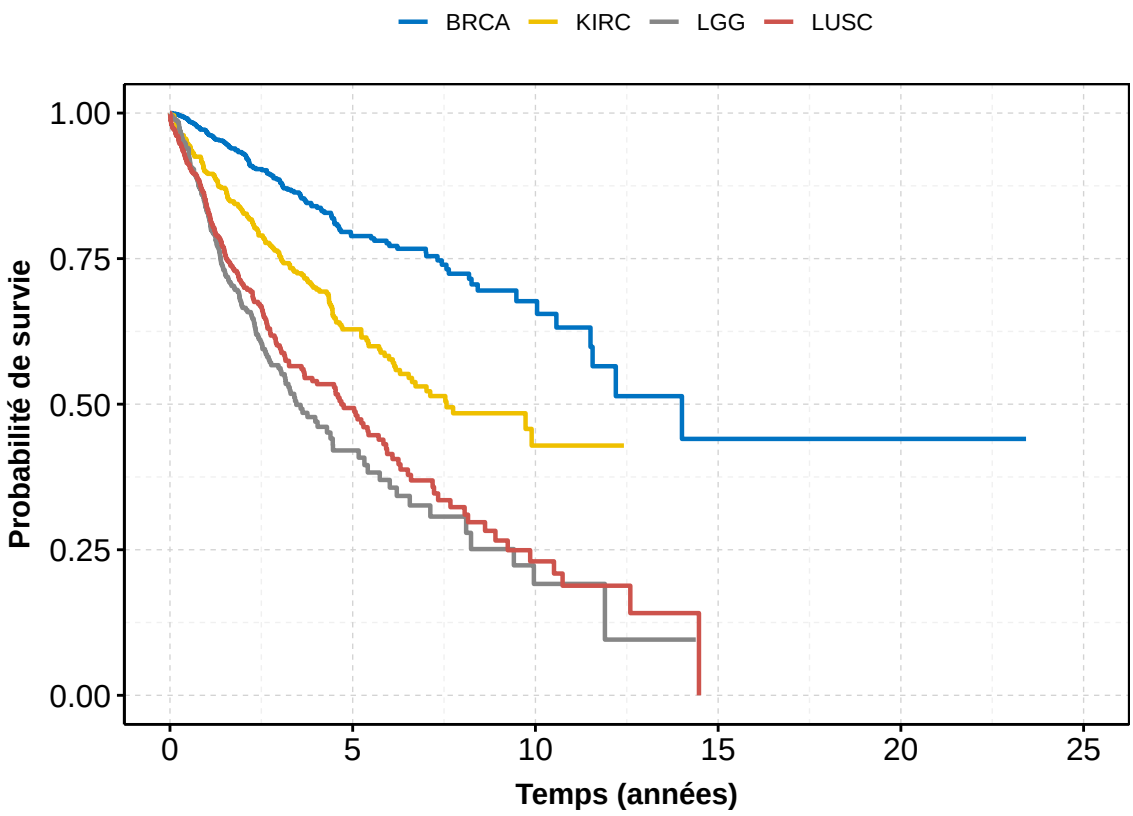


FIGURE 1.4 – **Courbes de survie obtenue à partir des données TCGA pour quatre cancers.** Selon les recommandations de LIU et collab. [2018b], la survie globale est utilisée pour KIRC et LUSC, et la survie sans récidive est utilisée pour BRCA et LGG.

1.6.4 Approche semi-paramétrique : le modèle de Cox

Le modèle de Cox [COX, 1972] est très largement utilisé en médecine pour lier des covariables cliniques ou génétiques à des données de survie à travers la fonction de risque

(*hazard function*) h, définie à chaque instant t par :

$$h(t) = \lim_{h \to 0} \frac{\mathrm{P}(t \leq \mathrm{T} < t + h | \mathrm{T} \geq t)}{h}.$$

Cette fonction de risque représente une probabilité d'observer l'évènement à l'instant t sachant qu'il n'a pas été observé avant t (risque instantanée), et se modélise comme suit dans le modèle de Cox :

$$\begin{aligned} h(t; \mathrm{X}_1, \ldots, \mathrm{X}_p) &= h_0(t) \exp(\beta_1 \mathrm{X}_1 + \ldots + \beta_p \mathrm{X}_p) \\ &= h_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}), \end{aligned}$$

avec h_0 la fonction de risque de base, $\mathbf{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_p)^{\mathrm{T}}$ le vecteur contenant les covariables (*e.g.* niveau d'expression des gènes), et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ le vecteur de coefficients associés aux covariables. La fonction h_0 est commune à tous les patients et donne la forme générale à la fonction de risque h. Les covariables forment ainsi une combinaison linéaire dans l'exponentielle, mais sont reliés à la fonction de risque par une fonction non linéaire. On parle dans ce cas de modèle linéaire généralisé.

Le modèle de Cox est aussi appelé « modèle à risques proportionnels » car le rapport des fonctions de risque de deux patients j et k qui ont pour covariables $\mathbf{X}^{(j)}$ et $\mathbf{X}^{(k)}$ est constant au cours du temps :

$$\frac{h(t; \mathbf{X}^{(j)})}{h(t; \mathbf{X}^{(k)})} = \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^{(j)})}{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^{(k)})},$$

avec t le temps.

Pour un patient i avec des données de survie (t_i, δ_i) tel que $\delta_i = 1$, la quantité $\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^i)}{\sum_{l \in \mathrm{R}_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^l)}$ correspond à la probabilité que l'individu i subisse effectivement l'événement en t_i sachant qu'un événement a eu lieu en t_i. Le vecteur de coefficients $\boldsymbol{\beta}$ du modèle de Cox peut ainsi être estimé en maximisant la pseudo-vraisemblance proposée par Breslow [1972] et définie comme le produit de ces probabilités conditionnelles :

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1 \ldots n \text{ t.q. } \delta_i = 1} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^i)}{\sum_{l \in \mathrm{R}_i} \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}^l)},$$

avec R_i l'ensemble des patients à risque au temps t_i (ensemble des patients qui ont un temps de suivi strictement supérieur à t_i), et $i = 1, \ldots, n$ tel que $\delta_i = 1$.

Cette fonction est appelée « pseudo-vraisemblance » car elle est définie comme un produit de probabilités conditionnelles, et non comme un produit de fonctions de densité. Pour faciliter l'optimisation et le calcul numérique, la log-pseudo-vraisemblance l (et non la pseudo-vraisemblance) est utilisée pour estimer $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}(l(\boldsymbol{\beta})),$$

avec $l(\boldsymbol{\beta}) = \log(\mathrm{L}(\boldsymbol{\beta}))$ la log-pseudo-vraisemblance.

Il est important de noter que le modèle de Cox n'est pas intuitif dans le sens où le lien entre les variables explicatives (*i.e.* niveau d'expression des gènes) et le temps de survie (variable à expliquer) est défini de manière indirecte par le biais de la fonction de risque h. Cependant, la pseudo-vraisemblance ainsi définie permet de prendre efficacement en compte les censures, tout en s'affranchissant de modéliser, d'estimer, et d'émettre des hypothèses sur la fonction de risque de base h_0. On dit ainsi que le modèle de Cox est « semi-paramétrique ». Finalement, cette modélisation aboutit à un problème d'optimisation convexe, et le vecteur $\boldsymbol{\beta}$ peut ainsi être estimé par des procédures classiques. Dans cette étude, nous utilisons le package *glmnet* [FRIEDMAN et collab., 2010; SIMON et collab., 2011] où un algorithme de descente de coordonnées cycliques (*cyclical coordinate descent*) est implémenté pour estimer $\boldsymbol{\beta}$.

1.7 Métriques d'évaluation de la qualité de prédiction

1.7.1 L'indice pronostique

L'un des objectifs premiers du modèle de Cox est d'assigner un score de risque à un patient à partir de variables explicatives (*e.g.* données génétiques, variables cliniques, etc.). Pour cela, le vecteur $\boldsymbol{\beta}$ du modèle de Cox est estimé sur un jeu de données d'apprentissage contenant des patients pour lesquels à la fois les variables explicatives et les données de survie sont connues. Le score de risque pour le patient étudié, qui ne fait pas parti du jeu de données d'apprentissage et dont on ne connaît *a priori* pas les données de survie, est appelé indice pronostique - *prognostic index* - (PI). Il se définit comme suit :

$$\hat{\mathrm{PI}} = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X},$$

avec $\hat{\boldsymbol{\beta}}$ l'estimateur du vecteur $\boldsymbol{\beta}$ dans le modèle de Cox calculé avec les patients du jeu de données d'apprentissage, et $\mathbf{X}$ le vecteur contenant les variables explicatives du patient étudié.

La fonction de risque définie à la partie 1.6.4 apparaît donc comme une fonction croissante de l'indice pronostique (Fig. 1.5) :

$$h(t) = h_0(t)\exp(\mathrm{PI}).$$

A un instant t donné, plus l'indice pronostique est élevé, plus la fonction de risque est élevée (Fig. 1.5), et donc plus le pronostic est mauvais. Réciproquement, l'indice pronostique est une fonction croissante du risque. Tout comme les paramètres β_i, il peut être

positif comme négatif.

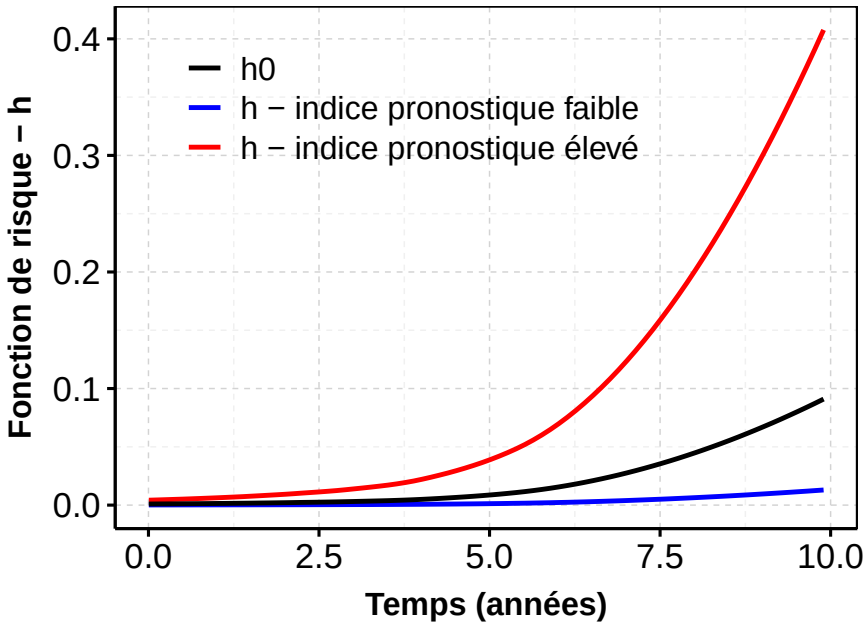


FIGURE 1.5 – **Fonction de risque pour différents patients.**
Fonction de risque de base h_0 (en noir), fonctions de risque pour un patient avec un indice pronostique élevé (en rouge), ou avec un indice pronostique faible (en bleu), estimées à partir des données d'ARN messagers sur le cancer du rein à cellules claires (KIRC). L'évènement étudié est le décès des patients (survie globale).

1.7.2 Procédure d'évaluation des modèles prédictifs

Le patient étudié (resp. le groupe de patients étudiés), souvent appelé « patient test » (resp. « jeu de données test », par opposition à « jeu de données d'apprentissage »), n'est pas inclus dans le jeu de données d'apprentissage afin d'évaluer le modèle. Nous assignons un indice pronostique au patient test (resp. à chaque patient du jeu de données test) à partir du vecteur $\hat{\boldsymbol{\beta}}$ estimé sur le jeu de données d'apprentissage, et cela permet d'évaluer si ce score de risque concorde bien avec les données de survie. De même, nous faisons le pré-filtrage sur le jeu de données d'apprentissage.

Classiquement, dans le contexte de l'étude des capacités de prédiction d'un modèle de survie [BOULESTEIX et collab., 2020; HERRMANN et collab., 2020; MILANEZ-ALMEIDA et collab., 2020], le calcul des indices pronostiques est fait en quatre étapes (Fig. 1.6.A) :

1. le jeu de données est subdivisé en un jeu de données d'apprentissage (typiquement 80% des patients) et un jeu de données test (les 20% des patients restants).
2. le vecteur $\boldsymbol{\beta}$ du modèle de Cox est estimé sur le jeu de données d'apprentissage.
3. les indices pronostiques sont calculés pour l'ensemble des patients du jeu de données test à partir du vecteur $\hat{\boldsymbol{\beta}}$ estimé à l'étape précédente.

4. une métrique d'évaluation des capacités de prédiction est calculée.

Pour augmenter la robustesse des résultats, nous répétons les trois étapes décrites précédemment 50 fois en utilisant 10 répétitions d'une validation croisée (Fig. 1.6.B) : le jeu de données est séparé en 5 échantillons (ou *folds*) de manière aléatoire, l'un de ces 5 échantillons définit le jeu de données test, et les 4 autres permettent de construire le jeu de données d'apprentissage. Nous répétons ce processus jusqu'à ce que les 5 échantillons aient été utilisé une fois comme jeu de données test. Ainsi, 5 métriques d'évaluation peuvent être calculées pour chacune des 10 validations croisées.

Nous nous attacherons à décrire les trois métriques d'évaluation de la qualité de prédiction les plus utilisées dans la littérature (*i.e.* C-index, score de Brier, et p-valeur du modèle de Cox univarié) dans les paragraphes suivants.

1.7.3 La concordance

L'indice de concordance (ou concordance, ou C-index) permet d'évaluer la discrimination d'un score de risque (capacité à bien classer les patients en terme de risque) en quantifiant la proportion de couples de patients dont les scores de risque sont en bon accord avec leurs données de survie. Pour deux patients i et j avec des indices pronostiques $\mathrm{PI}^{(i)}$, $\mathrm{PI}^{(j)}$, et des temps de survie $\mathrm{T}^{(i)}, \mathrm{T}^{(j)}$ respectivement, le C-index se définit comme :

$$\text{C-index} = \mathrm{P}(\mathrm{T}^{(i)} < \mathrm{T}^{(j)} | \mathrm{PI}^{(i)} > \mathrm{PI}^{(j)})$$

Deux patients sont dits « comparables » si le plus petit temps de suivi correspond à un évènement (*e.g.* décès, récidive), et non à une censure. En effet, dans ce dernier cas, il est impossible de déterminer si l'évènement du patient correspondant est intervenu avant ou après l'évènement du second patient.

Ainsi, deux patients comparables sont dits « concordants » si l'évènement du patient qui a le PI le plus grand a lieu avant l'évènement du patient qui a le PI le plus petit. Le C-index est donc une mesure de discrimination du modèle. Un C-index de 1 indique une parfaite concordance entre score de risque et survie. Dans ce cas, les patients peuvent être parfaitement classés en terme de risque associé à la survie. En revanche, un C-index de 0,5 correspond à une classification aléatoire des patients, et le modèle n'est donc aucunement utile en pratique.

De nombreux estimateurs de la concordance existent [Gönen et Heller, 2005; Uno et collab., 2011]. Dans cette étude, nous utiliserons l'estimateur de la concordance décrit par Harrell et collab. [1996] et théorisé par Pencina et D'Agostino [2004] :

$$\widehat{\text{C-index}} = \frac{1}{\#\mathrm{U}} \sum_{(i,j)\in \mathrm{U}} 1_{\{(t^{(i)}-t^{(j)})(\mathrm{PI}^{(i)}-\mathrm{PI}^{(j)})<0\}},$$

avec U l'ensemble des couples de patients comparables (le plus petit temps de suivi correspond à un évènement et non à une censure), #U la taille de U, 1_{A} la fonction indicatrice

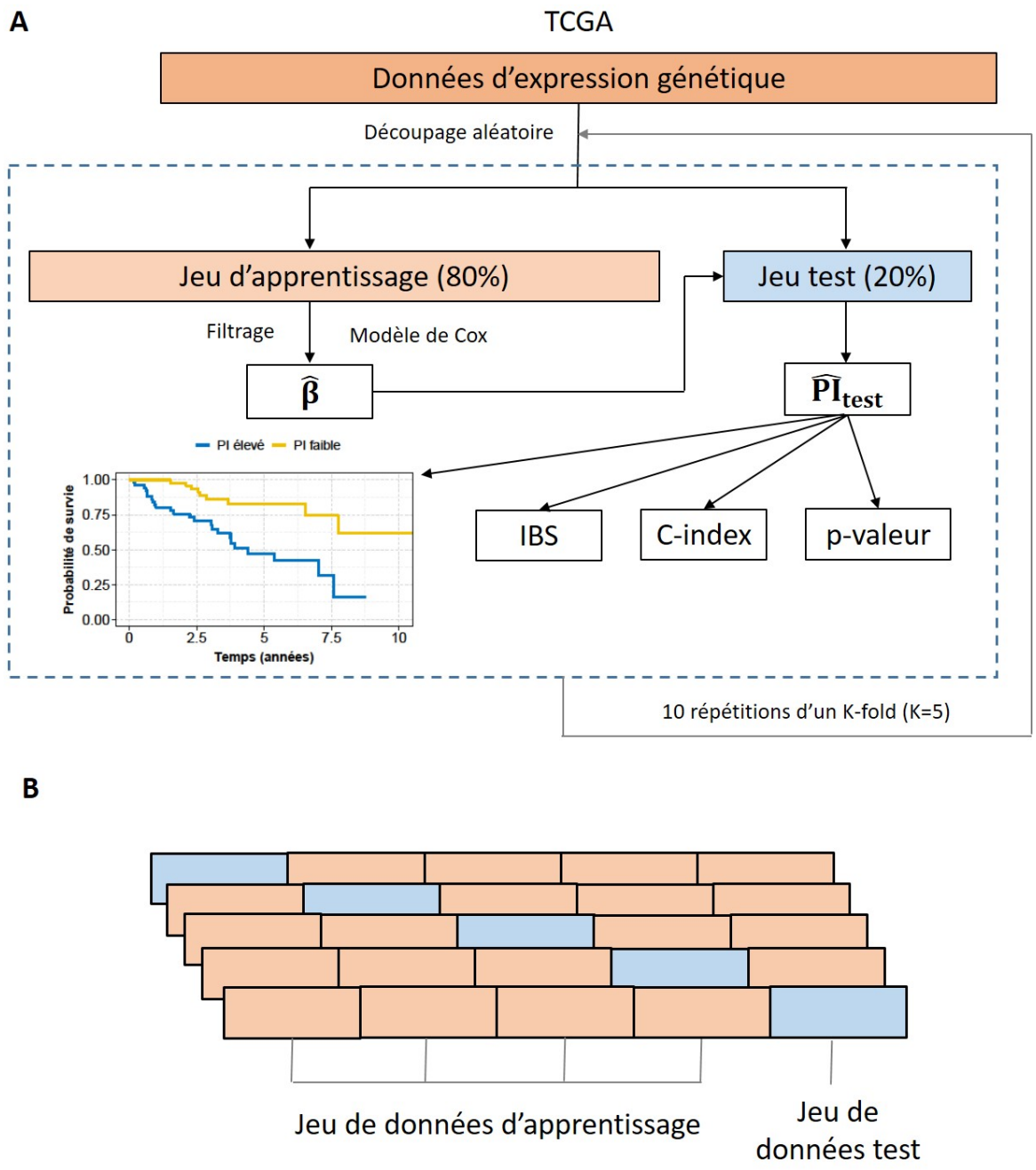


FIGURE 1.6 – **Schéma récapitulatif de l'évaluation des modèles de prédiction.**
(A) Nous entraînons un modèle de Cox sur 80% des patients, et nous calculons l'estimateur $\hat{\boldsymbol{\beta}}$ du modèle. Ce vecteur permet d'estimer les indices pronostiques $\widehat{\mathbf{PI}}_{\mathbf{test}}$ du jeu de données test, et nous calculons les différents indicateurs d'évaluation de la qualité de prédiction (*i.e.* score de brier intégré - IBS, C-index, p-valeur du modèle de Cox univarié, et estimateurs de Kaplan-Meier pour les patients avec un PI supérieurs à la médiane (bleu) et un PI inférieurs à la médiane (jaune)). Nous répétons ce processus 50 fois (10 répétitions d'une validation croisée, avec K = 5).
(B) Validation croisée avec K = 5 : nous séparons le jeu de données en 5 échantillons de manière aléatoire, l'un de ces 5 échantillons définit le jeu de données test, et les 4 autres permettent de construire le jeu de données d'apprentissage.

valant 1 si son argument A est vrai et 0 sinon, $t^{(i)}$ et $t^{(j)}$ les temps de suivi des patients i et j respectivement, $\text{PI}^{(i)}$ et $\text{PI}^{(j)}$ les indices pronostiques des patients i et j respectivement. Cet estimateur calcule donc la proportion de patients concordants parmi les patients comparables.

La fonction `concordance.index` du package `survcomp` [SCHRÖDER et collab., 2011] permet de calculer l'estimateur de la concordance proposé par HARRELL et collab. [1996].

1.7.4 La p-valeur du modèle de Cox univarié

La p-valeur du modèle de Cox univarié permet de tester l'association entre les indices pronostiques du jeux de données test et les temps de survie. Le calcul de cette métrique consiste à apprendre un modèle de Cox univarié sur le jeu de données test où l'unique variable explicative est l'indice pronostique. L'hypothèse nulle du test est « $\text{H}_0 : \beta = 0$ », où β est le coefficient du modèle de Cox univarié : $h(t;\text{PI}) = h_0(t)\exp(\beta \times \text{PI})$.

Le test du rapport de vraisemblance [BUSE, 1982] est préféré au test de Wald [KALBFLEISCH et PRENTICE, 2011] car il possède de meilleures propriétés lorsque la taille de l'échantillon est faible [KENDALL, 1946]. La statistique du test du rapport de vraisemblance s'exprime comme la différence de la log-vraisemblance du modèle après estimation du coefficient β et la log-vraisemblance du modèle nul : $\lambda_{\text{LR}} = -2\big(l(\hat{\beta} - l(0)\big)$. Asymptotiquement, cette statistique suit une loi du χ^2 à un degré de liberté. Ainsi, la p-valeur du modèle de Cox univarié permet de tester la plausibilité d'obtenir les indices pronostiques du jeu de données test sous l'hypothèse nulle qu'aucun lien n'existe entre les indices pronostiques (et donc le niveau d'expression des gènes) et la survie (*i.e.* $\beta = 0$). Elle peut donc être interprétée comme une mesure du niveau de corrélation entre les indices pronostiques et la survie, mesuré à travers un modèle linéaire généralisé.

La fonction `coxph` du package `survival` [THERNEAU, 2020] permet d'apprendre un modèle de Cox avec l'indice pronostique comme seule variable explicative de la survie, et de tester son association avec la survie par un test du rapport de vraisemblance.

1.7.5 Le score de Brier

La fonction de survie S peut-être estimée pour chaque patient i à partir de l'estimateur $\hat{\boldsymbol{\beta}}$ du modèle de Cox [COX et OAKES, 1984] et de l'estimateur du risque de base h_0. Le score de Brier - *Brier Score* - (BS) permet d'évaluer la précision de l'estimation de la fonction de survie à un instant t [GERDS et SCHUMACHER, 2006]. Il mesure la distance moyenne entre le statut et la fonction de survie prédite. En l'absence de censure et à un instant $t > 0$, le score de Brier se définit comme suit :

$$\text{BS}(t) = \frac{1}{n}\sum_{i=1}^{n}\left(1_{\{t^{(i)}>t\}} - \hat{\text{S}}(t,\text{PI}^{(i)})\right)^2,$$

avec $\hat{\text{S}}(t,\text{PI}^{(i)})$ l'estimateur de la fonction de survie à l'instant $t > 0$ pour le patient i.

Comme toute erreur quadratique, le score de Brier peut se décomposer en un terme de biais et un terme de variance :

$$(1_{\{t^{(i)}>t\}} - \hat{S}(t, PI^{(i)}))^2 = (1_{\{t^{(i)}>t\}} - S(t, PI^{(i)}))^2 + (S(t, PI^{(i)}) - \hat{S}(t, PI^{(i)}))^2,$$

avec $i = 1, ..., n$. Le premier terme de l'équation ci-dessus mesure la discrimination du modèle, alors que le second terme mesure la calibration du modèle (*i.e.* écart quadratique moyen entre les fonctions de survie individuelles théoriques et estimées).

En présence de censure, le score de Brier est ajusté par la distribution des censures $G(t) = P(C > t)$:

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(0 - \hat{S}(t, PI^{(i)}))^2 1_{\{t^{(i)}\leq t, \delta_i=1\}}}{\hat{G}(t^{(i)})} - \frac{(1 - \hat{S}(t, PI^{(i)}))^2 1_{\{t^{(i)}>t\}}}{\hat{G}(t)}\right),$$

avec $\hat{G}$ l'estimateur de la distribution des censure.

Plus le score de Brier est proche de 0, plus les prédictions individuelles sont bonnes, et un modèle prédictif devient utile lorsque le score de Brier est inférieur à 0,25. En effet, si les probabilités de survie sont de 0,5 pour tous les patients (*i.e.* incapacité du modèle à déterminer si l'évènement sera observé ou non), le score de Brier est de 0,25.

Le score de Brier intégré - *Integrated Brier Score* - (IBS) permet de calculer une moyenne des scores de Brier, et est souvent utilisé en pratique (Fig. 1.7). Pour un instant t donné, il se définit comme l'intégrale des scores de Brier entre 0 et t, divisé par t :

$$IBS(t) = \frac{1}{t}\int_0^t BS(s)ds$$

Les horizons temporels en terme de mortalité et de récidive sont variables suivant les cancers. Ainsi, dans notre étude, pour s'affranchir du choix de l'instant t, nous définissons la borne supérieure de l'intégrale comme le plus long temps de suivi du jeu de donnée test.

La fonction `sbrier.score2proba` du package `survcomp` [SCHRÖDER et collab., 2011] permet de calculer le score de Brier intégré.

1.7.6 Remarques sur les scores d'évaluation des performances de prédiction

Ainsi, alors que la concordance permet de mesurer si le modèle classe correctement les patients en terme de risque (*i.e.* discrimination), et que la p-valeur du modèle de Cox quantifie l'intensité de la corrélation entre les indices pronostiques et la survie, le score de Brier mesure la précision moyenne de la prédiction individuelle de la survie. Il est important de noter que :

— le calcul du score de Brier nécessite l'estimation de la fonction survie à un instant t et pour un patient $i = 1, \ldots, n$, $\hat{S}(t, PI^{(i)})$, et donc l'estimation de la fonction de risque

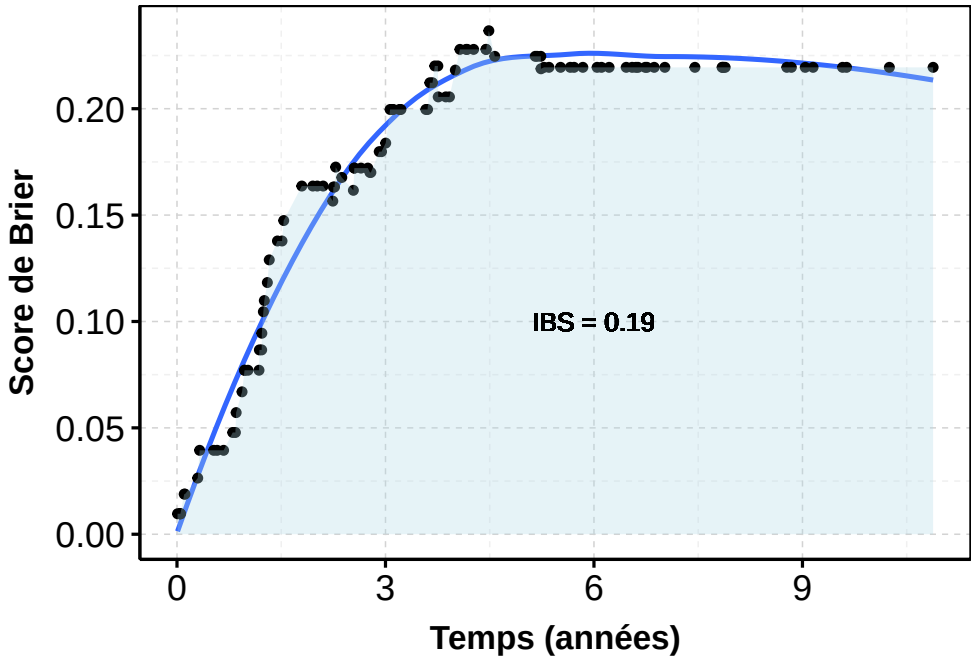


FIGURE 1.7 – **Score de Brier en fonction du temps calculé avec les données d'expression d'ARN messagers de TCGA pour le cancer du rein à cellules claires (KIRC).**
Le modèle de Cox est entrainé sur 80% des patients et le score de Brier est calculé sur les 20% restant. Les points noirs correspondent au score de Brier aux différents instants t, la courbe bleue est la courbe *loess* apprise sur les points noirs, et la partie bleue claire correspond au score de Brier intégré (IBS).

de base h_0. Cela n'est pas très intuitif dans le sens où le modèle de Cox a été construit pour s'affranchir de l'estimation de cette fonction.

— le score de Brier se compose en un score de calibration et de discrimination. Le score de Brier est ainsi une mesure des performances globales du modèle, mais demeure plus compliqué à interpréter que le C-index ou la p-valeur du modèle de Cox univarié.

— HARRELL et collab. [1996] suggère que re-calibrer un modèle est possible, alors que la discrimination ne peut être changé. En ce sens, la discrimination est plus importante que la calibration.

Malgré ces quelques remarques, le score de Brier demeure intéressant à étudier en pratique. Il permet en effet d'évaluer la capacité moyenne d'un modèle à prédire la survie à un horizon temporel t, ce que ne fait pas le C-index [BLANCHE et collab., 2019].

Ainsi, suivant l'objectif de l'étude :

— le C-index doit être choisi si la discrimination (*i.e.* capacité à bien classer les patients en terme de risque) est considérée.

— la p-valeur du modèle de Cox sera préféré si l'association avec la survie à travers un modèle de Cox est préféré.

— l'IBS doit être choisi si la prédiction individuelle de la survie est considérée.

1.8 La « malédiction de la grande dimension »

1.8.1 Définition

La « malédiction de la grande dimension » fait référence à l'apparition de phénomènes qui ont lieu lorsque le nombre p de variables excède le nombre n de patients. La dimension de l'espace dans lequel « vivent » les variables devient alors si grand que les données paraissent éparses et éloignées. Il devient alors difficile de dégager des généralités, et de nombreux algorithmes statistiques classiques sont mis à mal par un tel passage à l'échelle.

Dans le cadre des données d'expression génétiques, pour les sous-types de cancer les plus étudiés, la base de données TCGA contient typiquement 500 patients pour plus de 20 000 gènes. Dans ce contexte, les conséquences classiques de la malédiction de la grande dimension sont le manque de stabilité [JARDILLIER et collab., 2018] et le surajustement [PAVLOU et collab., 2015]. Dans ce cas, le modèle « colle » trop aux données d'apprentissage (*i.e.* les estimateurs du modèle changent radicalement lorsque les données d'apprentissage changent), et il est difficile de généraliser les résultats sur un nouveau jeu de données.

Deux stratégies distinctes permettent de remédier à ce fléau de la grande dimension, et elles visent toutes deux à réduire le nombre de variables considérées :

— le pré-filtrage univarié des gènes.

— les méthodes de pénalisation.

1.8.2 Pré-filtrage univarié des gènes

Le pré-filtrage univarié des gènes consiste à assigner un score à chaque gène individuellement (*e.g.* variabilité du gène, test d'association entre l'expression du gène et la survie), et à sélectionner uniquement les gènes dont le score est inférieur (ou supérieur) à un certain seuil. Cette approche a l'avantage d'être intuitive et simple mathématiquement, mais ne tient pas compte des corrélations entre les gènes [BØVELSTAD et collab., 2007]. Des variables très corrélées ont des scores proches et donc une forte probabilité d'être toutes sélectionnées, ce qui entraîne une redondance d'information. En revanche, de telles approches peuvent être intéressantes à considérer en tant que première étape de filtrage afin de supprimer les gènes *a priori* non-informatifs (*e.g.* très faible variabilité, faible association individuelle avec la survie), avant d'inclure uniquement les variables retenues dans un modèle multivarié. BENNER et collab. [2010] ont monté sur données simulées que le pré-filtrage peut permettre d'obtenir de meilleures prédictions avec le modèle de Cox, et l'étude de l'impact du pré-filtrage sur la prédiction fera l'objet du chapitre 3.

Les méthodes de pénalisation sont classiquement utilisées en grande dimension pour sélectionner un sous-ensemble de prédicteurs, et nous nous attacherons à les décrire dans la partie ci-dessous. Elles seront utilisées dans tous les chapitres de ce manuscrit.

1.8.3 La régression pénalisée

Les différentes formes classiques de pénalisation

L'idée sous-jacente des méthodes de pénalisation est de réduire la complexité de la solution en pénalisant les valeurs trop élevées des coefficients β_j, $j = 1,\dots,p$, du modèle de Cox afin de rétrécir vers 0 les coefficients des gènes les moins informatifs. On peut par exemple considérer que seul un sous-ensemble des 20000 gènes disponibles dans les jeux de données est réellement utile pour expliquer la survie et chercher un solution parcimonieuse (*i.e.* seuls quelques coefficients β_j sont non nuls). Les gènes sélectionnés sont alors définis comme les variables qui ont un coefficient β_j différent de 0. Dans cette étude, quatre méthodes de pénalisation sont utilisées : ridge [Verweij et Van Houwelingen, 1994], lasso [Tibshirani, 1997], elastic net [Zou et Hastie, 2005], et adaptive elastic net [Zou et Zhang, 2009]. Il est important de noter que contrairement aux autres méthodes, la pénalisation ridge ne fait pas de sélection, dans le sens où les coefficients β_j, $j = 1,\dots,p$, estimés sont rétrécis vers 0, mais restent en général non nuls.

Mathématiquement, l'idée des méthodes de pénalisation est d'ajouter une contrainte sur le vecteur de coefficients $\boldsymbol{\beta}$ lors de la maximisation de la pseudo-vraisemblance décrite à la partie 1.6.4. Pour chacune des méthodes, le programme d'optimisation à résoudre prend la forme suivante :

— Lasso

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(l(\boldsymbol{\beta}) - \lambda ||\boldsymbol{\beta}||_1 \right)$$

— Elastic Net (EN)

$$\hat{\boldsymbol{\beta}}(\text{EN}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(l(\boldsymbol{\beta}) - \lambda(\alpha||\boldsymbol{\beta}||_1 + \frac{1-\alpha}{2}||\boldsymbol{\beta}||_2^2) \right)$$

— Ridge

$$\hat{\boldsymbol{\beta}}(\text{ridge}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_2^2 \right)$$

— Adaptive Elastic Net (AEN) (procédure en deux étapes)

1. estimer le vecteur $\hat{\boldsymbol{\beta}^0}$ en maximisant l avec la régression ridge.
2. ajouter des poids dans la pénalisation elastic net avec les coefficients β_j^0, $j = 1,\dots,p$ calculés à l'étape 1 :

$$\hat{\boldsymbol{\beta}}(\text{AEN}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \left(l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \hat{w}_j \left(\alpha|\beta_j| + \tfrac{1-\alpha}{2}|\beta_j|^2 \right) \right),$$

avec $\hat{w}_j = 1/|\hat{\beta}_j^0|$, $j \in \{1, ..., p\}$.

La norme $l1$, utilisée dans les pénalisations ci-dessus à l'exception de ridge, permet d'assigner des coefficients β_j, $j = 1, \ldots, p$ à 0. Cette propriété est importante pour obtenir des scores de risque interprétable et pour des applications cliniques [KLAU et collab., 2018]. La régression ridge ne fait donc pas de sélection, mais a montré de bonnes capacités de prédiction en pratique [BØVELSTAD et collab., 2007]. En particulier, dans le cas de la régression linéaire et en présence de fortes corrélations dans les données, il a été observé que la pénalisation ridge obtient de meilleurs résultats de prédiction que lasso [TIBSHIRANI, 1996]. Ainsi, l'ajout d'une norme $l2$ dans l'elastic net permet de combiner les avantages des régressions lasso et ridge : (i) sélectionner un sous-ensemble de gènes pour la première, et (ii) obtenir de meilleures capacités de prédiction et de stabilité pour la deuxième.

La première étape de l'adaptive elastic net permet d'obtenir la propriété d'oracle [FAN et PENG, 2004] sous certaines conditions sur le jeu de données. Cette propriété stipule que lorsque la taille de l'échantillon tend vers l'infini ($n \to +\infty$), (i) la sensibilité tend vers 1 et le taux de faux positifs tend vers 0 (la méthode recouvre parfaitement les gènes qui ont un coefficient β_j, $j = 1, \ldots, p$ différent de 0), et (ii) les coefficients non nuls estimés sont normalement distribués, avec la même moyenne et le même écart-type que si la vérité terrain était connue par avance.

Choix du poids λ accordé à la pénalisation

Le paramètre λ présent dans le programme de maximisation permettant d'estimer les coefficients $\boldsymbol{\beta}$ pondère l'influence de la pénalité. En effet, plus λ est élevé, plus l'on accorde de l'importance à la pénalisation, et plus les coefficients $\boldsymbol{\beta}$ seront réduits vers 0. En pratique, le paramètre λ est fixé par validation croisée (K=5) par la procédure décrite ci-dessous :

1. partitionner l'ensemble d'apprentissage en 5 sous-ensembles.
2. estimer le vecteur $\boldsymbol{\beta}$ en maximisant la pseudo-vraisemblance avec pénalisation (partie 1.6.4) sur 4 sous-ensembles ($\boldsymbol{\beta}_{-i}$ dans la formule ci-dessous).
3. calculer une déviance à partir de la pseudo-log-vraisemblance totale l (les données des 5 sous-ensemble sont utilisées) et de la pseudo-log-vraisemblance l_{-i} calculée à partir des données des 4 sous-ensembles utilisés précédemment [HOUWELINGEN et collab., 2006] :
$$\hat{\text{CV}}_i(\lambda) = l(\boldsymbol{\beta}_{-i}(\lambda)) - l_{-i}(\boldsymbol{\beta}_{-i}(\lambda))$$
4. Recommencer ce procédé à partir de l'étape 3, en enlevant un autre ensemble i et en gardant la même valeur de λ.

   On obtient alors 5 valeurs pour $\hat{\text{CV}}_i(\lambda)$, ce qui nous permet de calculer une « dé-

viance moyenne » (ou « erreur moyenne ») pour le λ considéré :

$$\hat{\mathrm{CV}}(\lambda) = \frac{1}{5}\sum_{i=1}^{5} \hat{\mathrm{CV}}_i(\lambda)$$

Cette procédure est répétée pour différentes valeurs de λ en gardant le même partitionnement des patients. Deux heuristiques classiques sont utilisées pour fixer le poids de la pénalisation λ [SIMON et collab., 2011] (Fig. 1.8) :

— choisir le λ qui minimise l'erreur moyenne (« $\lambda_{\min}$ »).

— choisir le plus grand λ tel que l'erreur moyenne est inférieure à un écart-type du l'erreur moyenne minimale (« λ_{1se} »).

La deuxième méthode permet de fixer une valeur de λ plus importante et d'assigner plus de poids à la pénalité. Ainsi, moins de co-variables seront sélectionnées, et le modèle sera plus parcimonieux.

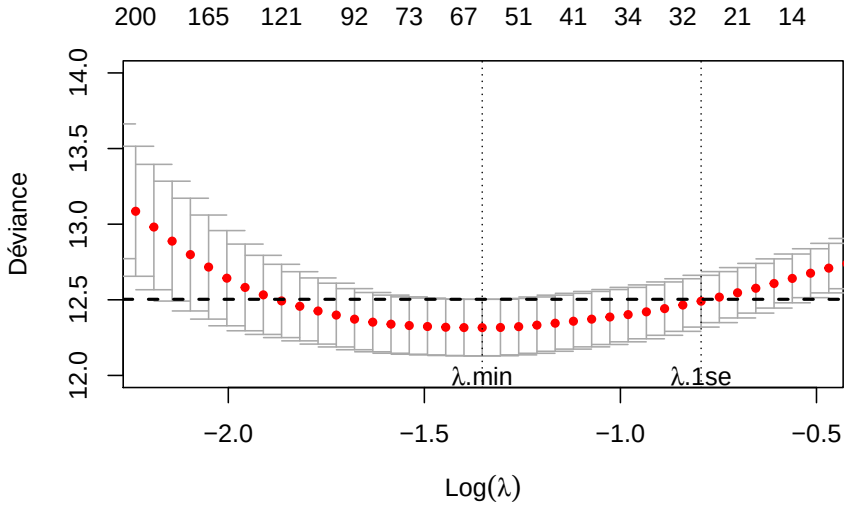


FIGURE 1.8 – **Calcul du poids de la pénalité elastic net qui minimise la déviance globale.**
Les données d'expression d'ARN messagers de TCGA pour le cancer du rein à cellules claires (KIRC) sont utilisées.
Chaque point rouge correspond à la moyenne de la déviance globale calculé par validation croisée (K=5, ordonnée) pour une séquence de λ donnée (abscisse, log(λ)) ; les barres d'erreur verticales grises correspondent à un écart-type de la déviance moyenne. La ligne verticale en pointillée la plus à gauche correspond à la valeur du λ qui minimise la déviance globale moyenne (« λ.min ») ; celle la plus à droite correspond au plus grand λ tel que l'erreur moyenne est inférieure à un écart-type du l'erreur moyenne minimale (ligne pointillée verticale, « λ.1se »). Le nombre de gènes sélectionnés est indiqué en haut du graphique pour quelques valeurs de λ.

Choix du poids α accordé à la norme $l1$ dans la pénalisation elastic net

Dans le paragraphe précédent consacré aux formes classiques de pénalisation, nous avons vu que les pénalisations elastic net et adaptive elastic net possède un hyperparamètre supplémentaire, α. Ce paramètre permet de répartir les poids assignés aux normes $l1$ et $l2$ dans la pénalisation. Plus α est élevé, plus le poids accordé à la norme $l1$ est élevé, et plus celui accordé à la norme $l2$ est faible. Ainsi, plus α est élevé, plus le nombre de gènes sélectionnés est faible.

Ce paramètre α peut-être fixé par l'utilisateur [JIANG et collab., 2016], ou fixé par validation croisée sur un vecteur de valeurs comprises entre 0 et 1. Dans ce dernier cas, tout comme pour le choix du poids λ accordé à la pénalisation, la déviance du modèle de Cox est classiquement retenue [MILANEZ-ALMEIDA et collab., 2020; OJEDA et collab., 2016]. Une double validation croisée menée à la fois sur λ et α est alors effectuée, et le processus de calcul est augmenté d'un facteur correspondant à la taille du vecteur des α testés.

Pour KIRC et avec la pénalisation elastic net, cette déviance reste relativement stable suivant les valeurs de α, et le nombre de gènes sélectionnés commence à se stabiliser à partir de 0,3 (Fig. 1.9). Cette tendance se retrouve pour l'ensemble des jeux de données étudiés (données non montrées). Ainsi, pour que les avantages de la pénalisation ridge soient pris significativement en compte, nous avons choisi α = 0,3 pour les pénalisations elastic net et adaptive elastic net pour l'ensemble des cancers.

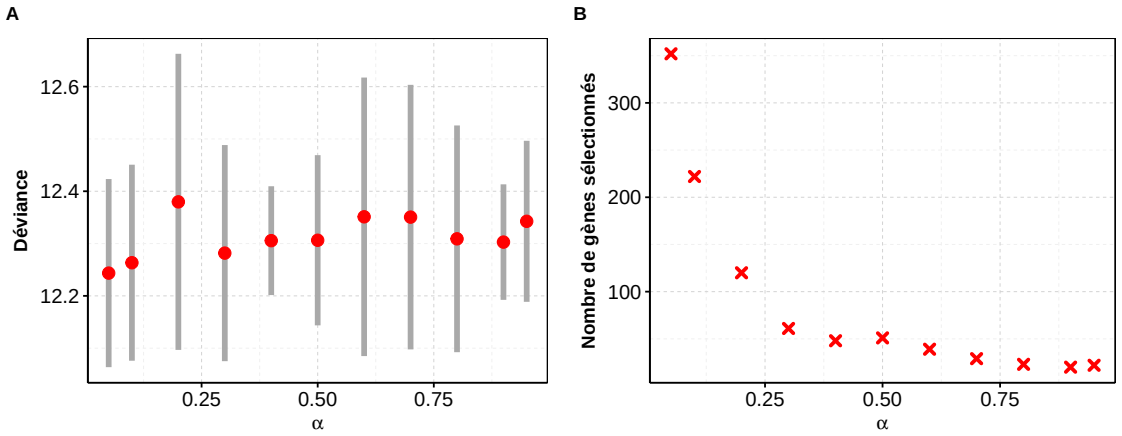


FIGURE 1.9 – **Déviance globale (A) et nombre de gènes sélectionnés (B) pour différentes valeurs de α dans la pénalisation elastic net pour les données mRNA-seq de KIRC.**
La séquence de α testée est 0,05; 0,1; 0,2; ...; 0,8; 0,9; 0,95, et correspond à la répartition du poids assigné à la norme $l1$ et $l2$.
(A) Déviance globale (points rouges) calculée par validation croisée (K=5) pour le λ qui minimise la déviance globale (λ.min) en fonction des différentes valeurs de α. Les barres d'erreur verticales grises correspondent à un écart-type de la déviance moyenne.
(B) Nombre de gènes sélectionnés en fonction de α.

1.9 Conclusions

1.9.1 Tableau récapitulatif des données de survie et mRNA-seq de TCGA

TABLEAU 1.2 – **Tableau récapitulatif des données de survie et mRNA-seq pour les 26 cancers de TCGA retenus.**
Nous avons calculé les C-index médians avec la pénalisation ridge par 10 répétitions d'une validation croisée (K=5). Les cancers sont classés par ordre décroissant de ces C-index médians.

Cancer	Nom	n patients	Taux de censure	Survie à 3 ans	C-index
ACC	Carcinome adrénocortical (glandes surrénales)	77	0,65	0,75	0,87
KIRP	Carcinome papillaire du rein	269	0,84	0,87	0,82
UVM	Mélanome de la choroïde	77	0,73	0,74	0,79
MESO	Mésothéliome (surfaces mésothéliales des poumons, de l'abdomen et du cœur)	85	0,14	0,19	0,74
LGG	Gliome cérébral de bas grade	510	0,62	0,56	0,73
KIRC	Carcinome du rein à cellules claires	526	0,67	0,76	0,72
CESC	Cancer épidermoïde du col utérin	288	0,76	0,72	0,71
PRAD	Adénocarcinome de la prostate	490	0,81	0,8	0,7
LIHC	Carcinome hépatocellulaire du foie	358	0,65	0,62	0,68
UCEC	Carcinome du corps utérin	541	0,83	0,83	0,67
BLCA	Carcinome urothélial (vessie)	371	0,56	0,48	0,65
LAML	Leucémie myéloïde aigüe (sang)	149	0,38	0,31	0,65
BRCA	Carcinome du sein infiltrant	1079	0,87	0,88	0,63
HNSC	Carcinome épidermoïde de la tête et du cou	491	0,57	0,57	0,63
LUAD	Adénocarcinome pulmonaire	488	0,63	0,61	0,63
THCA	Carcinome de la thyroïde	499	0,9	0,88	0,62
PAAD	Adénocarcinome du pancréas	175	0,47	0,34	0,61
COAD	Adénocarcinome du colon	432	0,78	0,78	0,61
THYM	Thymome (thymus)	119	0,82	0,85	0,61
GBM	Glioblastome multiforme (cerveau)	152	0,21	0,094	0,58
STAD	Adénocarcinome de l'estomac	385	0,59	0,47	0,57
OV	Cystadénocarcinome séreux des ovaires	300	0,39	0,61	0,55
LUSC	Carcinome épidermoïde du poumon	485	0,57	0,59	0,55
TGCT	Tumeurs des cellules germinales testiculaires	124	0,73	0,75	0,54
READ	Adénocarcinome du rectum	156	0,77	0,69	0,52
ESCA	Carcinome de l'œsophage	167	0,6	0,43	0,49

1.9.2 Résumé

Ainsi, malgré les efforts de recherche importants mis en œuvre pour une meilleure compréhension biologique du cancer, la recherche de nouveaux traitements, et une meilleure prise en charge des patients, de nombreux enjeux demeurent.

Tout d'abord, le « principe d'unicité du cancer » (partie 1.1.3) rend difficile la prise en charge et le suivi des patients. En effet, la diversité du microenvironnement tumoral, des caractéristiques génomiques des cellules tumorales, et des types cellulaires obligent à un suivi individualisé.

Ensuite, nous avons vu que les molécules d'ARN caractérisent le niveau d'expression d'un gène, et jouent un rôle important dans l'agressivité des cancers (partie 1.2.4). Certains gènes s'expriment anormalement (*i.e.* sur-expression ou sous expression par rapport aux gènes de tissus sains), et le niveau d'expression des gènes peut permettre de classer les patients en terme de risque associé à la survie. La technologie « RNA-seq » permet de mesurer ces niveaux d'expression (partie 1.3). Le coût du séquençage a fortement diminué au cours des dix dernières années, mais il demeure encore trop élevé pour une utilisation routinière en clinique. Le coût et la qualité du séquençage sont croissants de la profondeur de séquençage.

Le modèle de Cox permet de modéliser les données de survie et de les lier à des covariables explicatives (*e.g.* cliniques, génomiques). Les données RNA-seq comportent environ 20 000 gènes et typiquement 500 patients (partie 1.5.1). Le nombre de variables explicatives ($\sim$ 20 000) dépasse donc largement le nombre d'échantillons ($\sim$ 500), et ce cas est appelé la « malédiction de la grande dimension ». Les données paraissent alors éparses et éloignées, et les propriétés mathématiques classiques sont mises à mal. Deux méthodes permettent de remédier à ce passage à l'échelle : le pré-filtrage univarié des gènes, et les méthodes de pénalisation de la pseudo-vraisemblance.

1.9.3 Objectifs de la thèse

Dans ce contexte, les objectifs de cette thèse sont :

1. de comparer les différentes formes classiques de pénalisation du modèle de Cox (*i.e.* ridge, lasso, elastic net, adaptive elastic net) sur les données réelles de TCGA et données simulées (chapitre 2).
2. d'étudier l'impact du pré-filtrage univarié des gènes sur les prédictions de la survie obtenues avec le modèle de Cox pénalisé (chapitre 3).
3. d'étudier l'impact de la profondeur de séquençage sur les prédictions obtenues avec le modèle de Cox pénalisé dans le but d'optimiser les prédictions et le nombre d'échantillons séquencés sous contrainte de coûts (chapitre 4).

Dans ce manuscrit, nous utiliserons le modèle de Cox pénalisé pour prédire la survie ou la récidive des patients et ainsi améliorer leur suivi et leur prise en charge en cli-

nique (partie 1.1.4). Notons cependant que ces modèles peuvent être utilisé pour d'autres tâches, comme la recherche de marqueurs génétiques prédictifs de la réponse au traitements [TERNÈS et collab., 2017]. Nous utiliserons les données provenant de la base de données TCGA (partie 1.5.1).

1.9.4 Contexte de la thèse

L'analyse de données mRNA-seq de tumeurs dans des modèles de survie est une nouvelle thématique au sein du laboratoire BCI (Biologie du Cancer et de l'Infection). L'objectif de cette thèse était donc aussi de développer des compétences internes au laboratoire, et de lancer de nouveaux projets de recherche autour de cette thématique.

J'ai effectué l'ensemble des analyses présentées dans ce manuscrit, mais ce travail demeure collectif car les retours, les discussions, et les suggestions de mes encadrants ont largement contribué à mon travail de thèse. Ainsi, j'emploierai le pronom « nous » dans la suite du manuscrit.

1.10 Références

ALBERTS, B. et collab.. 2018, *Essential Cell Biology*, W. W. Norton Company, Oxford, UK, ISBN 978-0-393-69109-2. 6

BARTEL, D. P. 2018, «Metazoan MicroRNAs», *Cell*, vol. 173, nº 1, doi :10.1016/j.cell.2018.03.006, p. 20–51, ISSN 00928674. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418302861. 7

BEER, D. G. et collab.. 2002, «Gene-expression profiles predict survival of patients with lung adenocarcinoma», *Nature Medicine*, vol. 8, nº 8, doi :10.1038/nm733, p. 816–824, ISSN 1078-8956. 8

BENJAMINI, Y. et Y. HOCHBERG. 1995, «Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, nº 1, p. 289–300, ISSN 0035-9246. URL https://www.jstor.org/stable/2346101, publisher : [Royal Statistical Society, Wiley]. 16

BENJAMINI, Y. et T. P. SPEED. 2012, «Summarizing and correcting the GC content bias in high-throughput sequencing», *Nucleic Acids Research*, vol. 40, nº 10, doi : 10.1093/nar/gks001, p. e72–e72, ISSN 0305-1048. URL https://academic.oup.com/nar/article/40/10/e72/2411059, publisher : Oxford Academic. 10

BENNER, A. et collab.. 2010, «High-dimensional Cox models : the choice of penalty as part of the model building process», *Biometrical Journal. Biometrische Zeitschrift*, vol. 52, nº 1, doi :10.1002/bimj.200900064, p. 50–69, ISSN 1521-4036. 28

BERGER, A. C. et collab.. 2018, «A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers», *Cancer Cell*, vol. 33, nº 4, p. 690–705. 4

BLANCHE, P. et collab.. 2019, «The c-index is not proper for the evaluation of t-year predicted risks», *Biostatistics (Oxford, England)*, vol. 20, nº 2, doi :10.1093/biostatistics/kxy006, p. 347–357, ISSN 1468-4357. 27

BOULESTEIX, A.-L. et collab.. 2020, «Statistical learning approaches in the genetic epidemiology of complex diseases», *Human Genetics*, vol. 139, nº 1, doi :10.1007/s00439-019-01996-9, p. 73–84, ISSN 1432-1203. URL https://doi.org/10.1007/s00439-019-01996-9. 22

BRAY, F., J. FERLAY, I. SOERJOMATARAM, R. L. SIEGEL, L. A. TORRE et A. JEMAL. 2018, «Global cancer statistics 2018 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries», *CA Cancer J Clin*, vol. 68, nº 6, p. 394–424. 3

BRESLOW, N. 1972, «Contribution to the Discussion of the Paper by D.R. Cox», *Journal of the Royal Statistical Society B*, vol. 34, p. 2016–2017. 20

BUSE, A. 1982, «The Likelihood Ratio, Wald, and Lagrange Multiplier Tests : An Expository Note», *The American Statistician*, vol. 36, nº 3, doi :10.2307/2683166, p. 153, ISSN 00031305. URL https://www.jstor.org/stable/2683166?origin=crossref. 25

BØVELSTAD, H. M. et collab.. 2007, «Predicting survival from microarray data–a comparative study», *Bioinformatics (Oxford, England)*, vol. 23, nº 16, doi :10.1093/bioinformatics/btm305, p. 2080–2087, ISSN 1367-4811. 28, 30

CALIN, G. A. et C. M. CROCE. 2006, «MicroRNA signatures in human cancers», *Nat. Rev. Cancer*, vol. 6, nº 11, p. 857–866. 7

CHU, A., G. ROBERTSON, D. BROOKS, A. J. MUNGALL, I. BIROL, R. COOPE, Y. MA, S. JONES et M. A. MARRA. 2016, «Large-scale profiling of microRNAs for The Cancer Genome Atlas», *Nucleic Acids Research*, vol. 44, nº 1, doi :10.1093/nar/gkv808, p. e3, ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4705681/. 8

CIEŚLIK, M. et A. M. CHINNAIYAN. 2018, «Cancer transcriptome profiling at the juncture of clinical translation», *Nature Reviews Genetics*, vol. 19, nº 2, doi :10.1038/nrg.2017.96, p. 93–109, ISSN 1471-0064. URL https://www.nature.com/articles/nrg.2017.96, number : 2 Publisher : Nature Publishing Group. 12

COX, D. et D. OAKES. 1984, *Analysis of survival data.*, New York : Chapman and Hall/CRC. 25

COX, D. R. 1972, «Regression Models and Life-Tables», *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, vol. 34, nº 2, doi :10.1007/

978-1-4612-4380-9_37, p. 187–220, ISSN 00359246. URL http://link.springer.com/10.1007/978-1-4612-4380-9_37. 19

DUMBRAVA, E. I. et F. MERIC-BERNSTAM. 2018, «Personalized cancer therapy—leveraging a knowledge base for clinical decision-making», *Cold Spring Harbor Molecular Case Studies*, vol. 4, nº 2, doi :10.1101/mcs.a001578, ISSN 2373-2873. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5880252/. 7, 15

FAN, J. et H. PENG. 2004, «Nonconcave penalized likelihood with a diverging number of parameters», *Ann. Statist.*, vol. 32, nº 3, doi :10.1214/009053604000000256, p. 928–961. URL https://doi.org/10.1214/009053604000000256. 30

FODOR, S. P. et collab.. 1991, «Light-directed, spatially addressable parallel chemical synthesis», *Science*, vol. 251, nº 4995, doi :10.1126/science.1990438, p. 767–773, ISSN 0036-8075, 1095-9203. URL https://science.sciencemag.org/content/251/4995/767, publisher : American Association for the Advancement of Science Section : Research Articles. 8

FRIEDMAN, J. et collab.. 2010, «Regularization paths for generalized linear models via coordinate descent», *Journal of Statistical Software*, vol. 33, nº 1, p. 1–22. URL http://www.jstatsoft.org/v33/i01/. 21

GERDS, T. A. et M. SCHUMACHER. 2006, «Consistent estimation of the expected Brier score in general survival models with right-censored event times», *Biom J*, vol. 48, nº 6, p. 1029–1040. 25

GERLINGER, M. et collab.. 2012, «Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing», *New England Journal of Medicine*, vol. 366, nº 10, doi :10.1056/NEJMoa1113205, p. 883–892, ISSN 0028-4793. URL https://doi.org/10.1056/NEJMoa1113205, publisher : Massachusetts Medical Society _eprint : https ://doi.org/10.1056/NEJMoa1113205. 4

GUI, J. et H. LI. 2005, «Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data», *Bioinformatics (Oxford, England)*, vol. 21, nº 13, doi :10.1093/bioinformatics/bti422, p. 3001–3008, ISSN 1367-4803. 8

GÖNEN, M. et G. HELLER. 2005, «Concordance probability and discriminatory power in proportional hazards regression», *Biometrika*, vol. 92, nº 4, doi :10.1093/biomet/92.4.965, p. 965–970, ISSN 0006-3444. URL https://academic.oup.com/biomet/article/92/4/965/389449, publisher : Oxford Academic. 23

HAGERTY, R. G. et collab.. 2005, «Communicating prognosis in cancer care : a systematic review of the literature», *Annals of Oncology : Official Journal of the European Society*

for Medical Oncology, vol. 16, nº 7, doi :10.1093/annonc/mdi211, p. 1005–1053, ISSN 0923-7534. 5

HARRELL, F. E. et collab.. 1996, «Multivariable Prognostic Models : Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors», *Statistics in Medicine*, vol. 15, nº 4, doi :10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4, p. 361–387, ISSN 1097-0258. URL https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4. 23, 25, 27

HERRMANN, M. et collab.. 2020, «Large-scale benchmark study of survival prediction methods using multi-omics data», *Briefings in Bioinformatics*, doi :10.1093/bib/bbaa167, p. bbaa167, ISSN 1467-5463, 1477-4054. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa167/5895463. 22

HOOD, L. et S. H. FRIEND. 2011, «Predictive, personalized, preventive, participatory (P4) cancer medicine», *Nature Reviews Clinical Oncology*, vol. 8, nº 3, doi :10.1038/nrclinonc.2010.227, p. 184–187, ISSN 1759-4774. URL http://dx.doi.org/10.1038/nrclinonc.2010.227. 5

HOUWELINGEN, H. C. V. et collab.. 2006, «Cross-validated Cox regression on microarray gene expression data», *Statistics in Medicine*, vol. 25, nº 18, doi :10.1002/sim.2353, p. 3201–3216, ISSN 1097-0258. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2353, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2353. 30

JARDILLIER, R. et collab.. 2018, «Bioinformatics Methods to Select Prognostic Biomarker Genes from Large Scale Datasets : A Review», *Biotechnology Journal*, vol. 13, doi :10.1002/biot.201800103, p. 1–12. 28

JIANG, Y. et collab.. 2016, «Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis», *Genomics*, vol. 107, nº 6, doi :10.1016/j.ygeno.2016.04.005, p. 223–230, ISSN 1089-8646. 32

JUNTTILA, M. R. et F. J. DE SAUVAGE. 2013, «Influence of tumour micro-environment heterogeneity on therapeutic response», *Nature*, vol. 501, nº 7467, p. 346–354. 4

KALBFLEISCH, J. D. et R. L. PRENTICE. 2011, *The Statistical Analysis of Failure Time Data*, AMLBook, ISBN 9781118032985, doi :10.1002/9781118032985. 17, 25

KAPLAN, E. L. et P. MEIER. 1958, «Nonparametric estimation from incomplete observations», *Journal of the American Statistical Association*, vol. 53, nº 282, p. 457–481, ISSN 01621459. URL http://www.jstor.org/stable/2281868. 18

KENDALL, M. G. 1946, «The advanced theory of statistics.», *The advanced theory of statistics.*, , nº 2nd Ed. URL https://www.cabdirect.org/cabdirect/abstract/19471601829, publisher : Charles Griffin and Co., Ltd., London. 25

KLAU, S. et collab.. 2018, «Priority-Lasso : a simple hierarchical approach to the prediction of clinical outcome using multi-omics data», *BMC Bioinformatics*, vol. 19, nº 1, p. 322. 30

KUMAR-SINHA, C. et A. M. CHINNAIYAN. 2018, «Precision oncology in the age of integrative genomics», *Nature Biotechnology*, vol. 36, nº 1, doi :10.1038/nbt.4017, p. 46–60, ISSN 1546-1696. URL https://www.nature.com/articles/nbt.4017, number : 1 Publisher : Nature Publishing Group. 12

LAG, R. et collab.. 2008, «Seer cancer statistics review, 1975-2005, national cancer institute», . 3

LANDER, E. S. et M. S. WATERMAN. 1988, «Genomic mapping by fingerprinting random clones : a mathematical analysis», *Genomics*, vol. 2, nº 3, doi :10.1016/0888-7543(88)90007-9, p. 231–239, ISSN 0888-7543. 12

LIU, J. et collab.. 2018a, «An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics», *Cell*, vol. 173, nº 2, p. 400–416. 5

LIU, J. et collab.. 2018b, «An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics», *Cell*, vol. 173, nº 2, doi :10.1016/j.cell.2018.02.052, p. 400–416.e11, ISSN 00928674. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418302290. 15, 16, 19

LOOMANS-KROPP, H. A. et A. UMAR. 2019, «Cancer prevention and screening : the next step in the era of precision medicine», *NPJ Precis Oncol*, vol. 3, p. 3. 4

MAMMEDOV, T. et collab.. 2008, «A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates», *Computational biology and chemistry*, vol. 32, nº 6, doi : 10.1016/j.compbiolchem.2008.07.021, p. 452–457, ISSN 1476-9271. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727727/. 10

MANTEL, N. 1966, «Evaluation of survival data and two new rank order statistics arising in its consideration», *Cancer Chemother Rep*, vol. 50, nº 3, p. 163–170. 19

MEYER, M. et M. KIRCHER. 2010, «Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing», *Cold Spring Harbor Protocols*, vol. 2010, nº 6, doi :10.1101/pdb.prot5448, p. pdb.prot5448–pdb.prot5448, ISSN 1559-6095. URL http://www.cshprotocols.org/cgi/doi/10.1101/pdb.prot5448. 8

MILANEZ-ALMEIDA, P. et collab.. 2020, «Cancer prognosis with shallow tumor RNA sequencing», *Nature Medicine*, vol. 26, nº 2, doi :10.1038/s41591-019-0729-3, p. 188–192, ISSN 1078-8956, 1546-170X. URL http://www.nature.com/articles/s41591-019-0729-3. 12, 22, 32

OGINO, S., C. S. FUCHS et E. GIOVANNUCCI. 2012, «How many molecular subtypes? Implications of the unique tumor principle in personalized medicine», *Expert Rev. Mol. Diagn.*, vol. 12, nº 6, p. 621–628. 4

OJEDA, F. M. et collab.. 2016, «Comparison of Cox Model Methods in A Low-dimensional Setting with Few Events», *Genomics, Proteomics & Bioinformatics*, vol. 14, nº 4, doi :10.1016/j.gpb.2016.03.006, p. 235–243, ISSN 1672-0229. URL http://www.sciencedirect.com/science/article/pii/S1672022916300390. 32

PAPAVRAMIDOU, N. et collab.. 2010, «Ancient Greek and Greco–Roman Methods in Modern Surgical Treatment of Cancer», *Annals of Surgical Oncology*, vol. 17, nº 3, doi : 10.1245/s10434-009-0886-6, p. 665–667, ISSN 1068-9265. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820670/. 4

PAVLOU, M. et collab.. 2015, «How to develop a more accurate risk prediction model when there are few events», *BMJ*, vol. 351, doi :10.1136/bmj.h3868. URL https://www.bmj.com/content/351/bmj.h3868. 28

PENCINA, M. J. et R. B. D'AGOSTINO. 2004, «Overall C as a measure of discrimination in survival analysis : model specific population value and confidence interval estimation», *Statistics in Medicine*, vol. 23, nº 13, doi :10.1002/sim.1802, p. 2109–2123, ISSN 1097-0258. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1802, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1802. 23

PEZZELLA, F., M. TAVASSOLI et D. J. KERR. 2019, *Oxford Textbook of Cancer Biology*, Oxford University Press, Oxford, UK, ISBN 9780198779452. 4

RABIN, B. A. et collab.. 2013, «Predicting cancer prognosis using interactive online tools : a systematic review and implications for cancer care providers», *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, vol. 22, nº 10, doi : 10.1158/1055-9965.EPI-13-0513, p. 1645–1656, ISSN 1538-7755. 5

RAO, M. S. et collab.. 2019, «Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies», *Frontiers in Genetics*, vol. 9, doi :10.3389/fgene.2018.00636, ISSN 1664-8021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6349826/. 8

REUTER, J. A., D. V. SPACEK et M. P. SNYDER. 2015, «High-throughput sequencing technologies», *Mol. Cell*, vol. 58, nº 4, p. 586–597. 8

RISSO, D. et collab.. 2011, «GC-Content Normalization for RNA-Seq Data», *BMC Bioinformatics*, vol. 12, nº 1, doi :10.1186/1471-2105-12-480, p. 480, ISSN 1471-2105. URL https://doi.org/10.1186/1471-2105-12-480. 10

RITCHIE, M. E. et collab.. 2015, «limma powers differential expression analyses for RNA-sequencing and microarray studies», *Nucleic Acids Research*, vol. 43, nº 7, doi :10.1093/nar/gkv007, p. e47, ISSN 1362-4962. 14

ROBINSON, M. D. et A. OSHLACK. 2010, «A scaling normalization method for differential expression analysis of RNA-seq data», *Genome Biology*, vol. 11, nº 3, doi :10.1186/gb-2010-11-3-r25, p. R25, ISSN 1465-6906. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-3-r25. 13

ROBINSON, M. D. et collab.. 2010, «edgeR : a Bioconductor package for differential expression analysis of digital gene expression data», *Bioinformatics (Oxford, England)*, vol. 26, nº 1, doi :10.1093/bioinformatics/btp616, p. 139–140, ISSN 1367-4811. 13, 14

RUPAIMOOLE, R. et F. J. SLACK. 2017, «MicroRNA therapeutics : towards a new era for the management of cancer and other diseases», *Nature Reviews. Drug Discovery*, vol. 16, nº 3, doi :10.1038/nrd.2016.246, p. 203–222, ISSN 1474-1784. 7

SCHENA, M. et collab.. 1995, «Quantitative monitoring of gene expression patterns with a complementary DNA microarray», *Science (New York, N.Y.)*, vol. 270, nº 5235, doi :10.1126/science.270.5235.467, p. 467–470, ISSN 0036-8075. 8

SCHRÖDER, M. S. et collab.. 2011, «survcomp : an R/Bioconductor package for performance assessment and comparison of survival models», *Bioinformatics (Oxford, England)*, vol. 27, nº 22, doi :10.1093/bioinformatics/btr511, p. 3206–3208, ISSN 1367-4811. 25, 26

SENFT, D. et collab.. 2017, «Precision Oncology : The Road Ahead», *Trends in Molecular Medicine*, vol. 23, nº 10, doi :10.1016/j.molmed.2017.08.003, p. 874–898, ISSN 1471-4914. URL http://www.sciencedirect.com/science/article/pii/S1471491417301430. 12

SIMON, N., J. FRIEDMAN, T. HASTIE et R. TIBSHIRANI. 2011, «Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent», *Journal of Statistical Software*, vol. 39, nº 5, doi :10.18637/jss.v039.i05, ISSN 1548-7660. URL http://www.jstatsoft.org/v39/i05/. 21, 31

SIMS, D. et collab.. 2014, «Sequencing depth and coverage : key considerations in genomic analyses», *Nature Reviews Genetics*, vol. 15, nº 2, doi :10.1038/nrg3642, p. 121–132, ISSN 1471-0056, 1471-0064. URL http://www.nature.com/articles/nrg3642. 12

SORLIE, T. et collab.. 2001, «Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications», *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, nº 19, p. 10 869–10 874. 7

SUN, X.-X. et Q. YU. 2015, «Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment», *Acta Pharmacologica Sinica*, vol. 36, nº 10, doi :10.1038/aps.2015.92, p. 1219–1227, ISSN 1745-7254. URL https://doi.org/10.1038/aps.2015.92. 4

TERNÈS, N. et collab.. 2017, «Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces», *Biometrical Journal. Biometrische Zeitschrift*, vol. 59, nº 4, doi :10.1002/bimj.201500234, p. 685–701, ISSN 1521-4036. 35

THERNEAU, T. M. 2020, *A Package for Survival Analysis in R*. URL https://CRAN.R-project.org/package=survival, r package version 3.2-3. 18, 25

TIBSHIRANI, R. 1996, «Regression shrinkage and selection via the lasso», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, nº 1, p. 267–288, ISSN 00359246. URL http://www.jstor.org/stable/2346178. 30

TIBSHIRANI, R. 1997, «The lasso method for variable selection in the cox model», *Statistics in Medicine*, vol. 16, nº 4, doi :10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3, p. 385–395, ISSN 02776715. 29

UNO, H., T. CAI, M. J. PENCINA, R. B. D'AGOSTINO et L. J. WEI. 2011, «On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data», *Statistics in Medicine*, vol. 30, nº 10, doi :10.1002/sim.4154, p. 1105–1117, ISSN 1097-0258. 23

VERWEIJ, P. J. M. et H. C. VAN HOUWELINGEN. 1994, «Penalized likelihood in cox regression», *Statistics in Medicine*, vol. 13, nº 23-24, doi :10.1002/sim.4780132307, p. 2427–2436. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780132307. 29

VAN DE VIJVER, M. J. et collab.. 2002, «A gene-expression signature as a predictor of survival in breast cancer», *The New England Journal of Medicine*, vol. 347, nº 25, doi :10.1056/NEJMoa021967, p. 1999–2009, ISSN 1533-4406. 8

WANG, Z. et collab.. 2009, «RNA-Seq : a revolutionary tool for transcriptomics», *Nature Reviews Genetics*, vol. 10, nº 1, doi :10.1038/nrg2484, p. 57–63, ISSN 1471-0064. URL

https://www.nature.com/articles/nrg2484, number : 1 Publisher : Nature Publishing Group. 8

WHO. 2020, *WHO report on cancer : setting priorities, investing wisely and providing care for all*, World Health Organization. 3

ZOU, H. et T. HASTIE. 2005, «Regularization and variable selection via the elastic-net», *Journal of the Royal Statistical Society*, vol. 67, nº 2, doi :10.1111/j.1467-9868.2005.00503.x, p. 301–320, ISSN 13697412. 29

ZOU, H. et H. H. ZHANG. 2009, «On the adaptive elastic-net with a diverging number of parameters», *Annals of Statistics*, vol. 37, nº 4, doi :10.1214/08-AOS625, p. 1733–1751, ISSN 00905364. 29

Chapitre 2

Comparaison et évaluation des méthodes de pénalisation du modèle de Cox avec les données mRNA-seq

Sommaire

2.1 État de l'art et objectifs du chapitre 47

2.2 Comparaison des capacités de prédiction des méthodes de pénalisation du modèle de Cox sur données réelles 48

2.2.1 Choix du poids de la pénalisation 48

2.2.2 Comparaison des différentes métriques d'évaluation 50

2.2.3 Prédictions obtenues sur les différents jeux de données TCGA . . . 52

2.2.4 Nombre de gènes sélectionnés et temps de calcul 53

2.3 Combinaison des données cliniques et mRNA-seq pour prédire la survie 56

2.3.1 Bibliographie et objectifs 56

2.3.2 Méthodologie et données cliniques utilisées 56

2.3.3 Apport des données mRNA-seq pour prédire la survie 58

2.4 Procédure de simulation des données de survie et d'évaluation des pénalisations 60

2.4.1 Procédure de simulation 60

2.4.2 Exemple de données de survie simulées 62

2.4.3 Évaluation des méthodes de pénalisation sur données simulées . . 63

2.5 Capacités de prédiction sur données simulées 68

2.5.1 Influence des paramètres de simulation sur les prédictions oracles 68

2.5.2 Influence des paramètres de simulation sur les prédictions obtenues avec le modèle de Cox pénalisé 70

2.5.3 Comparaison des prédictions oracles et des prédictions obtenues par le modèle de Cox pénalisé 71

2.6 Capacités de sélection des méthodes de pénalisation du modèle de Cox . 74

2.6.1 Stabilité des gènes sélectionnés dans deux cohortes distinctes - données réelles . . . 74
2.6.2 Impact des paramètres de simulations sur les performances de sélection . . . 75
2.6.3 Influence du poids accordé à la pénalisation . . . 76
2.6.4 Comparaison des méthodes de pénalisation . . . 77
2.6.5 Influence du nombre de patients . . . 79
2.7 Influence des performances de sélection sur la prédiction . . . 80
2.8 Conclusions . . . 84
2.9 Références . . . 85

2.1 État de l'art et objectifs du chapitre

Le but de ce chapitre est de mener une analyse pan-cancer pour évaluer et comparer les pénalisations classiques du modèle de Cox (*i.e.* ridge, lasso, elastic net, adaptive elastic net, partie 1.8.3). Nous étudierons les 26 cancers de TCGA choisis dans la partie 1.5.2, et l'analyse portera à la fois sur les performances de prédiction et de sélection.

Tout d'abord, BØVELSTAD et collab. [2007] ont comparé les performances de prédiction des pénalisation ridge et lasso. De plus, leur étude porte sur les prédictions obtenues avec les N gènes qui ont la p-valeur du modèle de Cox univarié la plus faible, avec N fixé par validation croisée grâce au calcul de la déviance du modèle de Cox multivarié pour différentes valeurs de N. Trois jeux de données réelles sont utilisés (deux jeu de données de cancer du sein, et un jeu de données de cancer du sang), et les métriques d'évaluations choisies par les auteurs sont la p-valeur du test du log-rank (*i.e.* test de la différence entre les courbes de Kaplan-Meier des patients qui ont les indices pronostiques (PI) inférieurs et supérieurs à la médiane des PI), la p-valeur du modèle de Cox univarié (partie 1.7.4), et la différence entre la déviance du modèle avec prédicteurs et du modèle nul, sans prédicteur. Leurs résultats montrent que la pénalisation ridge obtient les meilleures prédictions sur les jeux de données choisis, et que l'utilisation des N gènes qui ont les p-valeurs du modèle de Cox les plus faible (avec N fixé par validation croisée) ne suffit pas à prédire correctement la survie.

Par la suite, BENNER et collab. [2010] ont comparé les pénalisations ridge, lasso, elastic net et adaptive lasso. Leur étude porte sur des données simulées, sur des données d'expression génétique provenant d'une cohorte de patients atteints d'une leucémie myéloïde aiguë (LAML, Tab. 1.2) [METZELER et collab., 2008], et sur des données d'expression génétique de patientes atteintes d'un cancer du sein (BRCA, Tab. 1.2) [VAN 'T VEER et collab., 2002]. Les métriques utilisées pour évaluer la sélection sur données simulées sont le nombre de faux positifs et le nombre de faux négatifs, définis respectivement comme le nombre de gènes sélectionnés à tort, et le nombre de gènes non sélectionnés à tort. La métrique utilisée pour évaluer les prédictions est le score de Brier intégré (partie 1.7.5). Comme conclusions pratiques, les auteurs suggèrent d'utiliser la pénalisation lasso ou elastic net, et de poursuivre les recherches sur le pré-filtrage des gènes avant l'utilisation du modèle de Cox. L'impact du pré-filtrage sur la prédiction fera l'objet du chapitre 3.

Ensuite, OJEDA et collab. [2016] ont montré que les pénalités ridge, lasso et elastic net dans le modèle de Cox ont des performances de prédiction équivalentes en grande dimension. Les auteurs ont utilisé des données simulées et des données de patients atteints d'une maladie de l'artère coronaire. Le C-index est utilisé pour évaluer les performances de prédiction des différents modèles.

Enfin, concernant les performances de sélection, ROBERTS et NOWAK [2014] ont montré que l'ensemble des gènes sélectionnés par la pénalisation lasso est instable et dépend de la validation croisée qui permet de calculer le poids de la pénalisation λ (partie 1.8.3).

Ainsi, l'objectif de ce chapitre est d'étendre les résultats décrits ci-dessus en menant une analyse pan-cancer des données mRNA-seq de TCGA et en utilisant toutes les métriques d'évaluation classiques des modèles. Pour la prédiction, nous utiliserons le C-index, l'IBS, et la p-valeur du modèle de Cox univarié (partie 1.7). Pour évaluer la sélection, nous avons choisi la proportion de faux positifs (*False Discovery Proportion* (FDP)) et la sensibilité. Enfin, les deux heuristiques classiques pour le choix du poids λ de la pénalisation (partie 1.8.3) seront étudiées.

Par souci de clarté, les figures contiendront les résultats obtenus pour les 10 cancers pour lesquels le C-index médian calculé avec les données mRNA-seq est le plus grand. Les figures obtenues pour les 16 autres cancers seront placées en annexe.

2.2 Comparaison des capacités de prédiction des méthodes de pénalisation du modèle de Cox sur données réelles

Pour évaluer les capacités de prédiction des différentes pénalisations du modèle de Cox, nous avons calculé 50 valeurs de C-index, IBS et p-valeur du modèle de Cox univarié par 10 répétitions d'une validation croisée (K=5) (partie 1.7.2). Pour chaque répétition, les quatre pénalités considérées sont appliquées (*i.e.* ridge, lasso, elastic net, adaptive elastic net).

2.2.1 Choix du poids de la pénalisation

Afin de déterminer quelle heuristique choisir pour le poids assigné à la pénalisation (*i.e.* $\lambda_{\min}$ ou λ_{1se}), nous avons fait un test des rangs signés de Wilcoxon entre les C-index obtenus avec $\lambda_{\min}$ et ceux obtenus avec λ_{1se} pour chacun des cancers. Nous avons effectué des tests similaires pour l'IBS et la p-valeur du modèle de Cox univarié. Pour certains jeux de données, aucun gène n'est sélectionné avec λ_{1se} et les métriques d'évaluation des prédictions (*i.e.* C-index, p-valeurs du modèle de Cox univarié et IBS) ne peuvent pas être calculées. Sur les 104 combinaisons possibles entre jeux de données (26 cancers) et méthodes (4), 14 comparaisons entre $\lambda_{\min}$ et λ_{1se} ne peuvent pas être effectuées. Ainsi, nous avons pu calculer 90 p-valeurs issues de tests des rangs signés de Wilcoxon et corrigées suivant les cancers pour chacune des méthodes.

Le premier test que nous avons effectué permet de déterminer si les prédictions obtenues avec $\lambda_{\min}$ sont meilleures que celles obtenues avec λ_{1se} (*i.e.* test des rangs signés de Wilcoxon unilatéral). Sur les 90 tests que nous avons effectués et après correction par Benjamini-Hochberg pour chaque méthode suivant les cancers, 41, 35 et 2 tests sont significatifs au niveau 0,05 pour le C-index, la p-valeur du modèle de Cox et l'IBS, respectivement. Plus précisément et pour le C-index, pour 16 cancers sur 26 les prédictions sont significativement meilleures avec $\lambda_{\min}$ pour ridge, 9 sur 19 pour lasso, 8 sur 19 pour elastic

net, et 8 sur 26 pour adaptive elastic net (Fig. 2.1, Fig. Annexe A.1 et A.2).

De la même manière, nous avons testé si les prédictions obtenus avec λ_{1se} sont meilleures que celles obtenues avec λ_{min} (*i.e.* test des rangs signés de Wilcoxon unilatéral). Sur les 90 tests que nous avons effectués et après correction par Benjamini-Hochberg pour chaque méthode suivant les cancers, respectivement 6, 5 et 35 tests sont significatifs au niveau 0,05 pour le C-index, la p-valeur du modèle de Cox et l'IBS, respectivement. Plus précisément, les C-index sont significativement meilleures avec λ_{1se} pour 6 cancers sur 26 pour adaptive elastic net, et pour aucun jeu de données pour les trois autres méthodes. Des résultats similaires sont obtenus pour la p-valeurs du modèle de Cox, avec 5 jeux de données pour lesquels les prédictions sont meilleurs avec λ_{1se} et la pénalisation ridge, et aucun pour les trois autres méthodes. Concernant l'IBS, les 35 tests significatifs se répartissent entre les pénalisations ridge (13 tests significatifs) et adaptive elastic net (22 test significatifs).

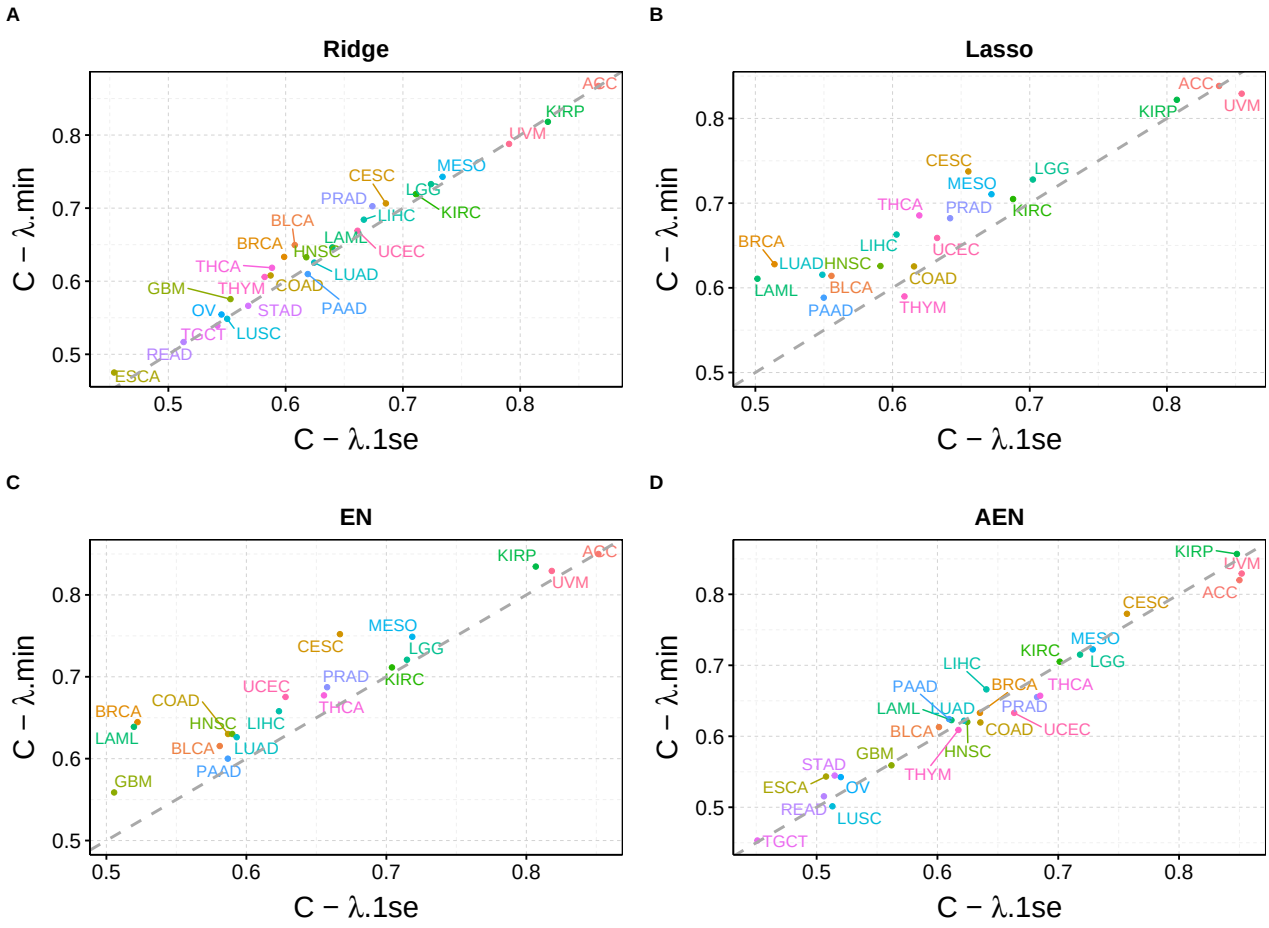


FIGURE 2.1 – **C-index médians obtenus avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés.**
Nous avons calculé les indices pronostiques par 10 répétitions d'une validation croisée (K=5). La médiane des 50 C-index est affichée ici pour chaque cancer.
(A) Ridge, (B) lasso, (C) elastic net (EN), (D) adaptive elastic net (AEN).

Ainsi, en prenant le C-index ou la p-valeur du modèle de Cox comme métriques d'éva-

luation, les prédictions sont significativement meilleures ou équivalentes avec λ_{min} pour ridge, lasso, et elastic net pour l'ensemble des jeux de données. Pour seulement 6 et 5 cancers pour le C-index et la p-valeur du modèle de Cox univarié, respectivement, λ_{1se} apporte de meilleures prédictions que λ_{min} pour adaptive elastic net. Ainsi, si l'on considère le C-index et la p-valeur du modèle de Cox comme mesure de prédiction, λ_{min} doit être préféré.

En revanche, il est important de noter que cette tendance est inversée lorsque l'IBS est défini comme métrique d'évaluation des prédictions pour ridge et adaptive elastic net. Ce résultat suggère que la calibration est meilleure avec λ_{1se} plutôt qu'avec λ_{min} pour ces deux méthodes. De plus, ASSEL et collab. [2017] ont montré que l'IBS est influencé par le sur-apprentissage. Nous pouvons ainsi envisager qu'une pénalisation plus importante permet de réduire le sur-apprentissage, et que λ_{1se} permet d'obtenir des scores de Brier plus faibles pour ridge et adaptive elastic net.

Finalement, au vu des résultats exposés dans cette partie et en considérant que la discrimination est plus importante que la calibration pour des applications pratiques HARRELL et collab. [1996], nous utiliserons systématiquement λ_{min} pour prédire la survie.

2.2.2 Comparaison des différentes métriques d'évaluation

Comme nous l'avons souligné en introduction, le C-index est une mesure de discrimination (*i.e.* capacité à classer les patients en termes de risque), la p-valeur du modèle de Cox univarié permet de quantifier l'intensité de la corrélation entre les indices pronostiques et la survie, et l'IBS mesure à la fois la discrimination et la calibration du modèle. Dans cette partie, afin d'obtenir une meilleure interprétation des résultats, nous allons analyser les corrélations qui existent entre les différentes métriques à travers l'ensemble des 26 cancers étudiés.

Nous observons que le C-index et la p-valeur du modèle de Cox univarié sont fortement corrélés (corrélations < -0,75 et p-valeurs d'un test de Pearson < 0,001 pour les quatre pénalisations, Tableau 2.1). La p-valeur du modèle de Cox univarié peut être vue comme une comme métrique d'évaluation des prédictions complémentaire du C-index : elle permet d'observer si, en plus d'une bonne capacité de discrimination, la valeur des indices pronostiques est corrélée à la survie à travers un modèle à risques proportionnels (linéaire généralisé). En effet, il est possible de changer la distribution des indices pronostiques sans changer le C-index (*i.e.* sans changer les rangs), mais en dégradant ou en améliorant la p-valeur du modèle de Cox univarié.

En revanche, les corrélations entre le C-index et l'IBS, et entre la p-valeur du modèle de Cox et l'IBS sont plus faible (-0,55 au plus et 0.45 au plus, respectivement). Ce résultat suggère que la calibration du modèle ne dépend pas de ses capacités de discrimination. En effet, si c'était le cas, l'IBS serait composé d'une somme de deux scores (partie 1.7.5) corrélés à la discrimination, et une corrélation plus marquée serait observée.

Pour certains cancers, nous observons que le C-index et l'IBS sont faibles (Fig. 2.2). Par exemple, pour GBM, le C-index médian obtenu avec la pénalisation elastic net est de 0,53 et l'IBS de 0,12. Cela suggère que malgré une incapacité à classer les patients en terme de risque, la calibration du modèle est bonne.

TABLEAU 2.1 – **Corrélations entre les trois métriques d'évaluation des prédictions.**
Nous calculons les corrélations pour chaque méthode suivant l'ensemble des 26 cancers étudiés (chiffres dans le tableau), et un test de Pearson est effectué (étoiles dans le tableau; n.s : non-significatif; + : p-valeur < 0,1; * : p-valeur < 0,05; ** : p-valeur < 0,01; *** : p-valeur < 0,001).

Méthode	C / p	C / IBS	p / IBS
ridge	-0,77 ***	-0,47 *	0,45 *
lasso	-0,85 ***	-0,53 **	0,42 *
EN	-0,79 ***	-0,55 **	0,35 +
AEN	-0,75 ***	-0,54 **	0,45 *

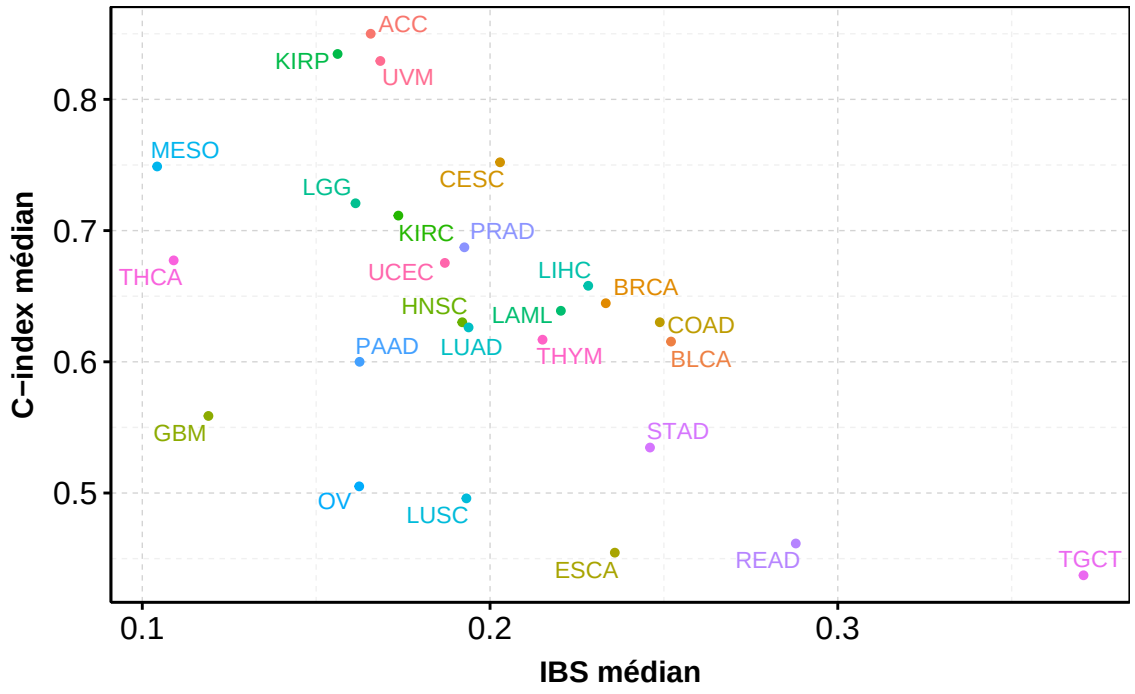


FIGURE 2.2 – **C-index médians en fonction des IBS médians calculés avec la pénalisation elastic pour l'ensemble des 26 cancers de TCGA étudiés.**
Nous avons calculé les indices pronostiques par 10 répétitions d'une validation croisée (K=5). La médiane des 50 valeurs est affichée ici.

2.2.3 Prédictions obtenues sur les différents jeux de données TCGA

Pour chaque cancer et chaque méthode, nous avons fait un test des rangs signés de Wilcoxon pour déterminer si les C-index obtenus avec la méthode étudiée sont significativement plus grands que ceux obtenus avec les trois autres méthodes. Nous considérerons que ce test est significatif si la p-valeur est inférieure à 0,05.

Tout d'abord, il est important de noter que la variance et les métriques d'évaluation des prédictions peuvent être influencées par le nombre de patients. Pour un jeu de données, plus le nombre de patients est important, plus les résultats seront stables, et plus la variance sera faible. De plus, les p-valeurs sont influencées par le taille des échantillons [THIESE et collab., 2016]. Ainsi, l'analyse que nous allons mener est pertinente pour comparer les méthodes au sein d'un même jeu de données, mais peut être biaisée par la taille des échantillons lorsque les résultats sont comparés entre les cancers.

Les résultats obtenus sont très dépendants du jeu de données étudié (Fig. 2.2, Fig. 2.3, et Fig. Annexe A.3). Par exemple, en prenant le C-index comme métrique d'évaluation, la pénalisation ridge obtient les meilleurs résultats pour LIHC, alors qu'elle est dominée par les trois autres méthodes (*i.e.* lasso, elastic net, adaptive elastic net) pour UVM (Fig. 2.3). Il apparaît ainsi difficile de tirer des conclusions générales et de préconiser une méthode plutôt qu'une autre pour l'ensemble des jeux de données.

Cependant, les C-index obtenus avec la pénalisation ridge sont significativement plus importants que ceux obtenus par les trois autres méthodes (*i.e.* lasso, elastic net, adaptive elastic net) pour 9 des 26 cancers étudiés (LGG, KIRC, PRAD, LIHC, BLCA, STAD, OV, LUSC et READ, p-valeurs < 0.05, tests des rangs signés de Wilcoxon unilatéraux avec correction de Benjamini-Hochberg, Fig. 2.3.A et Fig. Annexe A.3.A). La pénalité adaptive elastic net domine les autres méthodes pour KIRP, CESC et ESCA, et les meilleurs résultats sont obtenus avec le lasso pour THCA. Si la pénalité ridge semble la plus adaptée pour 9 des 26 cancers, elle est cependant dominée par les trois autres méthodes pour KIRP, UVM, CESC et THCA. Nous avons obtenu des résultats similaires en prenant la p-valeur du modèle de Cox univarié comme métrique d'évaluation : la pénalisation ridge domine les autres méthodes pour 6 cancers (ACC, MESO, LGG, KIRC, LIHC et BLCA), alors que le lasso obtient les meilleurs résultats pour THCA et adaptive elastic net pour CESC (Fig. 2.3.B et Fig. Annexe A.3.B). En prenant l'IBS comme métrique d'évaluation, ridge et lasso dominent les autres méthodes, chacune pour trois cancers (MESO, LGG, KIRC pour ridge, et UVM, THCA, COAD pour le lasso, Fig. 2.3.C et Fig. Annexe A.3.C). L'IBS apparaît donc moins sensible que les deux autres métriques (*i.e.* C-index et p-valeur du modèle de Cox univarié) pour comparer les méthodes.

Enfin, parmi les trois méthodes qui permettent de sélectionner un sous-ensemble de gènes et en choisissant le C-index comme critère de performance, elastic net domine lasso pour 10 des 26 cancers (MESO, LGG, KIRC, CESC, UCEC, BLCA, LAML, BRCA, PAAD, STAD), alors que de meilleurs résultats sont obtenus avec lasso pour seulement un cancer

(THCA). De même, elastic net obtient de meilleurs résultats qu'adaptive elastic net pour 7 cancers (ACC, MESO, LGG, KIRC, PRAD, UCEC, LAML), alors que l'inverse est vrai pour 5 cancers (KIRP, CESC, PAAD, OV, ESCA). Les résultats vont dans le même sens pour la p-valeur du modèle de Cox univarié et l'IBS. Ainsi, en moyenne, elastic net semble mieux prédire la survie que lasso et adaptive elastic net. Ce résultat reste cependant à nuancer car elastic net obtient de moins bons résultats qu'au moins une des deux autres méthodes pour certains jeux de données.

2.2.4 Nombre de gènes sélectionnés et temps de calcul

Le nombre de gènes sélectionnés est plus important pour elastic net et adaptive elastic net que pour lasso, à la fois pour λ_{min} et pour λ_{1se} (Fig. 2.4.A et Fig. Annexe A.4.A). En effet, la pénalisation lasso a tendance à ne sélectionner qu'une variable parmi un groupe de variables corrélées, alors que elastic net assigne des coefficients similaires, au signe près, aux variables d'un tel groupe [ZOU et HASTIE, 2005]. De plus, pour les 7 cancers qui ont les C-index médians les plus élevés (ACC, KIRP, UVM, MESO, LGG, KIRC, CESC), le nombre de gènes sélectionné par elastic net et adaptive elastic net est comparable, alors qu'il devient plus important pour adaptive elastic net pour les 19 autres cancers. La procédure en deux étapes permet de « forcer » l'algorithme à sélectionner des gènes.

Ensuite, le nombre médian de gènes sélectionnés avec λ_{1se} comme poids assigné à la pénalité est non nul pour ACC, KIRP, UVM, MESO, LGG et KIRC pour lasso, et devient nul pour les autres cancers (Fig. 2.4.A et Fig. Annexe A.4.A). Des résultats similaires sont observés pour elastic net, avec seulement deux jeux de données en plus de ceux évoqués ci-dessus pour lesquels le nombre médian de gènes sélectionnés est différent de 0 pour λ_{1se}. Ainsi, l'heuristique du λ_{1se} semble trop pénalisante, et ne permet pas de sélectionner un nombre suffisant de gènes dans le modèle.

Enfin, adaptive elastic net est une procédure en deux étapes (partie 1.8.3) et nécessite des temps de calculs plus importants que les trois autres méthodes (Fig. 2.4.B et Fig. Annexe A.4.B). En revanche, lasso, elastic net et ridge ont tous des temps de calcul similaires.

Remarquons que pour les cancers pour lesquels les prédictions obtenues sont les moins bonnes, nous observons un nombre important d'erreurs (*i.e.* erreur lors de l'appel de la fonction `glmnet` permettant d'apprendre un modèle de Cox pénalisé, ou aucun gène sélectionné) dans le calcul des métriques de prédiction (Fig. 2.4.C et Fig. Annexe A.4.C).

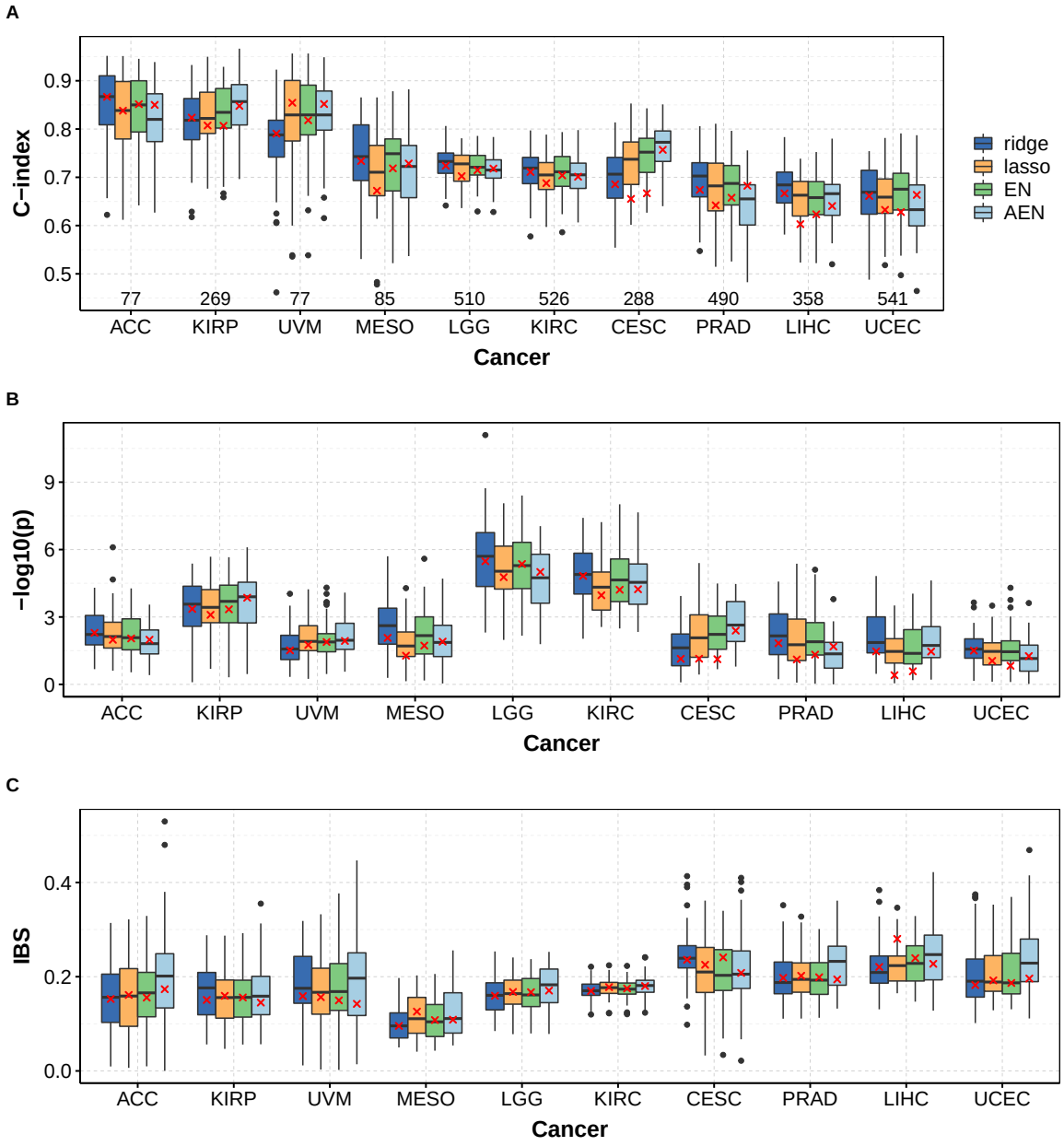


FIGURE 2.3 – **C-index (A), p-valeurs du modèle de Cox univarié (-log10) (B), et IBS (C) obtenus avec les données mRNA-seq.**

Pour chaque cancer et chaque méthode de pénalisation, les 50 C-index, p-valeurs, et IBS sont calculés par 10 répétitions d'une validation croisée (K=5) sur l'ensemble des gènes.

Les boites représentent les C-index (A), p-valeurs (-log10) (B), et IBS (C) obtenus avec λ_{min}, et les croix rouges représentent le C-index (A), p-valeurs (-log10) (B), et IBS (C) médians obtenus avec λ_{1se}.

Les nombres de patients dans les jeux de données sont notés en noir en bas du graphique A.

Les 10 cancers qui ont le plus grand C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 16 autres cancers étudiés sont placés en annexe (Fig. Annexe A.3).

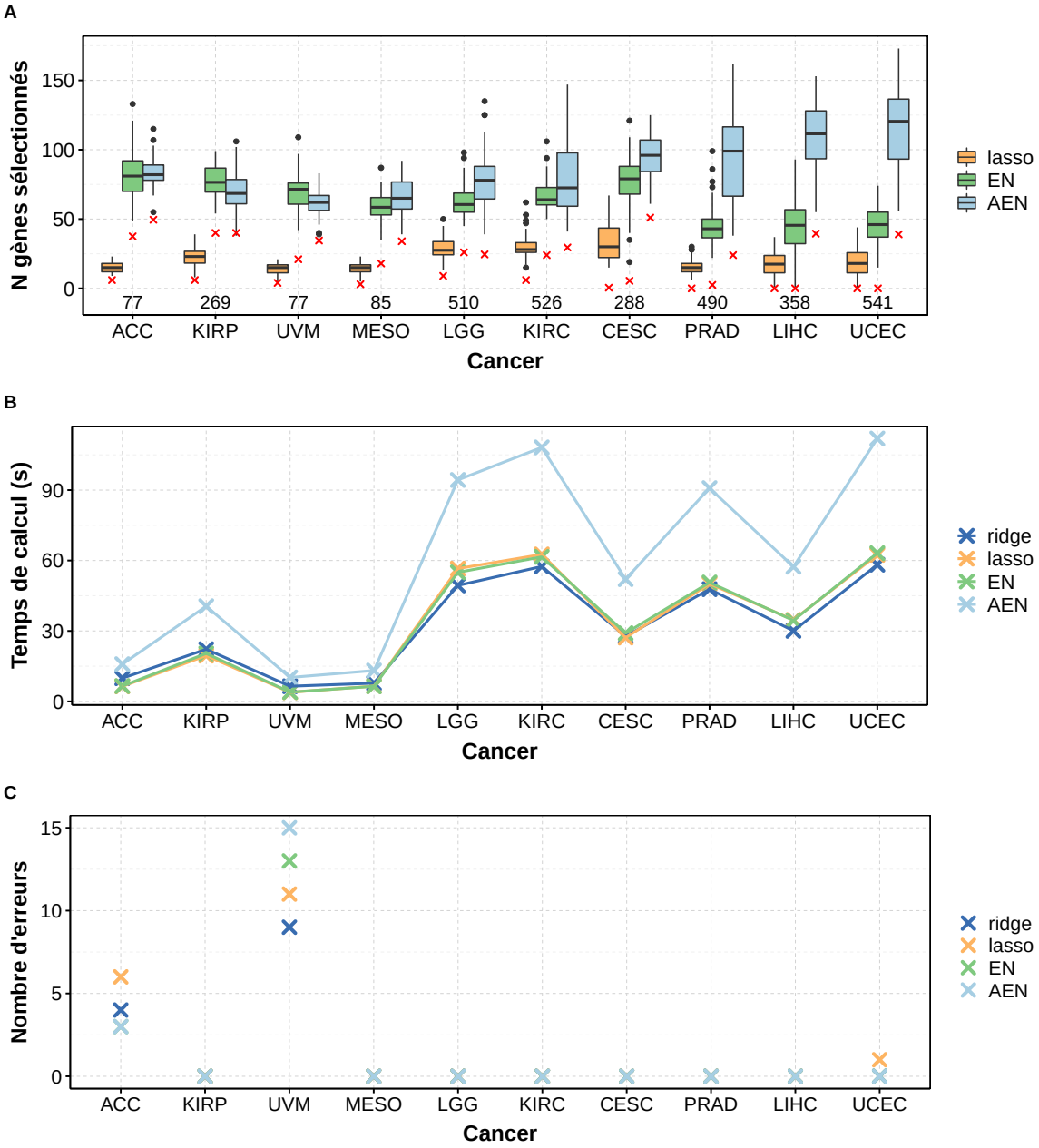


FIGURE 2.4 – **Nombre de gènes sélectionnés (A), temps de calcul médian (s) (B), et nombre d'erreurs dans le calcul du C-index avec λ_{min} (C) des méthodes de pénalisation obtenus avec les données mRNA-seq.**

Pour chaque cancer et chaque méthode de pénalisation, 50 modèles sont appris par 10 répétitions d'une validation croisée (K=5) sur l'ensemble des gènes. Pour chacun des 50 apprentissages, le nombre de gènes sélectionnés (A) et le temps de calcul (B) sont mesurés. Les croix rouges dans le graphique A représentent le nombre médian de gènes sélectionnés avec λ_{1se}. Une erreur correspond au cas où aucun gène n'est sélectionné, ou lorsqu'une erreur apparaît dans l'appel de la fonction `glmnet` permettant d'apprendre un modèle de Cox pénalisé.

Les nombres de patients dans les jeux de données sont notés en noir en bas du graphique A.

Les 10 cancers qui ont le plus grand C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 16 autres cancers étudiés sont placés en annexe (Fig. Annexe A.4).

2.3 Combinaison des données cliniques et mRNA-seq pour prédire la survie

2.3.1 Bibliographie et objectifs

Les données transcriptomiques ont montré une valeur ajoutée par rapport aux critères cliniques pour prédire la survie pour certains cancers. Par exemple, BØVELSTAD et collab. [2009] ont montré que la combinaison des données cliniques classiques et des données d'expression génétique permet d'obtenir de meilleures prédiction par rapport aux données cliniques seules. Ensuite, ZHAO et collab. [2015], dans un travail visant à étudier l'intérêt de mixer plusieurs types de données génomiques, ont montré que les données qui ont la plus forte valeur prédictive sont le couplage des données cliniques et d'expression génétique. Le C-index est utilisé comme métrique d'évaluation des prédictions, et quatre cancers de TCGA sont utilisés. Ces résultats ont été renforcé par l'étude de 14 cancers de TCGA [ZHU et collab., 2017], et les auteurs ont observé un intérêt des données génomiques pour la prédiction pour la moitié des cancers étudiés. Par la suite, MILANEZ-ALMEIDA et collab. [2020] ont montré que les données mRNA-seq apportent une valeur prédictive par rapport aux données cliniques seules sur l'ensemble des 13 jeux de données de TCGA étudiés, et en prenant le C-index comme métrique d'évaluation. L'intérêt des données génomiques a été montré dans de nombreux autres papiers [LÓPEZ DE MATURANA et collab., 2019].

Cependant, l'apport des données mRNA-seq pour prédire la survie n'est pas toujours vérifié. En effet, VOLKMANN et collab. [2019] ont montré que les données d'expression génétiques n'apportent aucune valeur prédictive pour le cancer du sein (BRCA) sur deux jeux de données indépendants lorsqu'un nombre suffisant de variables cliniques est utilisé. Les auteurs ont utilisé l'IBS et le C-index comme mesures de la qualité de la qualité de prédiction. De plus, dans une étude visant à intégrer les données multi-omiques pour prédire la survie, HERRMANN et collab. [2020] ont montré que les prédictions obtenues avec la meilleure méthode d'intégration testée sont seulement très légèrement supérieures à celles obtenues avec uniquement les données cliniques. Dans cette étude, les auteurs ont ajouté aux variables cliniques classiques des données spécifiques à chaque cancer étudié.

Dans ce contexte, l'objectif de cette partie et d'étudier l'apport des données mRNA-seq pour prédire la survie par rapport aux données cliniques classiques sur les 26 jeux de données de TCGA choisis à la partie 1.5.2.

2.3.2 Méthodologie et données cliniques utilisées

Certaines données cliniques sont classiquement utilisées en pratique pour établir un pronostique de la récidive ou de la survie des patients.

Le « grade » est déterminé lors de l'examen anatomopathologique (*i.e.* examen à l'oeil

nu et au microscope des tissus prélevés). Il se définit à partir de l'apparence des cellules cancéreuses (plus une cellule cancéreuse ressemble aux cellules normales, moins elle est agressive, et plus une cellule s'est modifiée par rapport aux cellules normales, plus elle est agressive), de la forme des noyaux (plus les noyaux des cellules cancéreuses sont gros, et de tailles et de formes variées, plus la tumeur est agressive), et du nombre de cellules en division (plus une cellule cancéreuse se développe vite, plus elle se divise rapidement, et plus le risque de propagation du cancer dans l'organisme augmente). Un score définissant le grade est attribué à la tumeur suivant ces trois critères. Le grade 1 correspond aux tumeurs les moins agressives, et le grade 4 aux plus agressives.

La classification internationale TNM permet de définir le « stade » d'un cancer. La lettre « T » est l'initiale de « *Tumor* » (tumeur) et correspond à la taille de la tumeur (T0, T1, T2, T3, T4) ; la lettre « N » est l'initiale de « *Node* » (ganglion) et indique si les ganglions lymphatiques ont été envahis (N1) ou non (N0) ; la lettre « M » est l'initiale de « *Metastasis* » (métastase) et indique la présence (M1) ou non (M0) de métastases. Ces trois variables permettent de définir le stade (stade 1, 2, 3, ou 4). Les définitions du stade diffèrent suivant les cancers en fonction de leur valeur pronostique [Brierley et collab., 2017], et nous utiliserons ces trois variables T, N et M comme prédicteurs dans le modèle de Cox.

Ces six données cliniques sont classiquement utilisées pour prédire la survie [Lu et collab., 2018; Milanez-Almeida et collab., 2020; Rosenberg et collab., 2005]. Nous les retenons uniquement lorsqu'elles sont présentes pour au moins 90% des patients de TCGA. L'âge est disponible pour les 26 cancers, le genre pour les 22 cancers non unisexe (*i.e.* CESC et UCEC sont des cancers touchant uniquement les femmes, PRAD uniquement les hommes, et le jeu de données TCGT ne contient que des patients masculins). Le grade n'est pas disponible pour 10 cancers, la variable T pour 19, la variable N pour 18, et la variable M pour 14 (Tab. 2.2).

La fonction `coxph` du package `survival` [Therneau, 2020] permet d'apprendre un modèle de Cox sans pénalisation avec un nombre restreint de variables. Les variables cliniques catégorielles (*i.e.* genre, grade, T, N, et M) sont subdivisées en différentes variables à l'intérieur de cette fonction. Par exemple, le grade est décrit par trois variables :

— « grade 2 » : 1 pour les patients de grade 2, 0 pour les autres.

— « grade 3 » : 1 pour les patients de grade 3, 0 pour les autres.

— « grade 4 » : 1 pour les patients de grade 4, 0 pour les autres.

Pour étudier la valeur ajoutée des ARNm pour la prédiction de la survie par rapport à ces variables cliniques classiques, nous avons calculé 50 C-index obtenus avec ces six variables cliniques (lorsqu'elles sont disponibles), 50 C-index obtenus avec les données mRNA-seq, et 50 C-index en mixant les données cliniques et les données mRNA-seq. Ces 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5) dans chacun des cas (Fig. 1.6).

Cancer	Age	Genre	Grade	T	N	M
ACC	1	1	0	1	1	0
BLCA	1	1	1	1	1	1
BRCA	1	1	0	1	1	1
CESC	1	0	1	0	0	0
COAD	1	1	0	1	1	1
ESCA	1	1	0	1	1	0
GBM	1	1	0	0	0	0
HNSC	1	1	1	1	1	0
KIRC	1	1	1	1	1	1
KIRP	1	1	0	1	1	1
LAML	1	1	0	0	0	0
LGG	1	1	1	0	0	0
LIHC	1	1	1	1	1	1

Cancer	Age	Genre	Grade	T	N	M
LUAD	1	1	0	1	1	1
LUSC	1	1	0	1	1	1
MESO	1	1	0	1	1	1
OV	1	1	1	0	0	0
PAAD	1	1	1	1	1	1
PRAD	1	0	0	1	0	0
READ	1	1	0	1	1	1
STAD	1	1	1	1	1	1
TGCT	1	0	0	1	1	0
THCA	1	1	0	1	1	1
THYM	1	1	0	0	0	0
UCEC	1	0	1	0	0	0
UVM	1	1	0	1	1	1

TABLEAU 2.2 – **Variables cliniques présentes pour chacun des 26 cancers étudiés.**
1 : la variable clinique est présente pour au moins 90% des patients ; 0 : la variable clinique n'est pas présente pour au moins 10% des patients.

Ensuite, pour mixer les données cliniques et les données mRNA-seq, les indices pronostiques calculés avec les données mRNA-seq sont utilisés comme une nouvelle covariable explicative, en plus des variables cliniques. L'objectif est que les variables cliniques gardent une grande importance et ne soient pas « noyées » sous le nombre de gènes. Pour évaluer la valeur ajoutée des données mRNA-seq dans la prédiction par rapport aux données cliniques seules, nous faisons un test de Wilcoxon unilatéral pour chaque cancer. Ce test permet d'observer si le C-index médian obtenu avec le mixte des données cliniques et mRNA-seq est significativement supérieur à celui obtenu avec les données cliniques seules.

Enfin, au vu des résultats de la partie 2.2.3, nous avons choisi d'utiliser la pénalisation ridge pour calculer les indices pronostiques du jeu de données mRNA-seq.

2.3.3 Apport des données mRNA-seq pour prédire la survie

Ainsi, pour 12 cancers (ACC, BLCA, BRCA, CESC, HNSC, KIRP, LGG, LIHC, MESO, PAAD, PRAD, THYM) sur les 26 étudiés, l'ajout des données mRNA-seq aux données cliniques améliorent le C-index significativement par rapport aux données cliniques seules (p-valeur $< 0{,}05$; test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg) (Fig. 2.5 et Fig. Annexe A.5). Dans la partie 2.3.1, nous avons vu que VOLKMANN et collab. [2019] n'ont pas observé de valeur ajoutée des données mRNA-seq dans la prédiction pour BRCA lorsqu'un nombre suffisant de variables cliniques est utilisé. Leurs résultats divergent ainsi des nôtres, mais cela s'explique par l'utilisation d'une variables clinique spécifique au cancer du sein dans leur étude, la présence ou non de récepteurs des œstrogènes. En revanche, pour COAD, ESCA, GBM, KIRC, LAML, LUAD, LUSC, OV, READ, STAD, TGCT, THCA, UCEC, et UVM, les données mRNA-seq n'apportent pas de valeur prédictive par

rapport aux données cliniques.

De manière intéressante, pour COAD et TGCT, les données mRNA-seq ne permettent pas d'obtenir de bonnes prédictions (C-index médian < 0.55), alors que les données cliniques permettent de prédire correctement la survie (C-index médian > 0.70).

En prenant la p-valeur du modèle de Cox univarié comme métrique d'évaluation des prédictions, l'ajout des données mRNA-seq aux données cliniques est pertinent pour 12 cancers (ACC, BLCA, CESC, HNSC, KIRC, KIRP, LGG, LIHC, MESO, PAAD, PRAD, UVM), dont 10 sont communs avec ceux cités au paragraphe précédent. Ces résultats concordent globalement avec ceux obtenus par MILANEZ-ALMEIDA et collab. [2020], sauf pour LAML, et LUAD. Différentes raisons peuvent expliquer ces divergences :

— Dans la validation croisée, MILANEZ-ALMEIDA et collab. [2020] ont utilisé 50% des patients à la fois pour le jeu d'apprentissage et le jeu test (80% et 20% dans notre cas, respectivement).

— Les tests effectués par MILANEZ-ALMEIDA et collab. [2020] portent sur les intervalles de confiance des p-valeurs du modèle de Cox univarié après correction par la méthode de Benjamini-Hochberg, alors que nous avons utilisé un test des rangs signés de Wilcoxon.

— Nous avons corrigé les p-valeurs obtenues pour chaque cancer par la méthode de Benjamini-Hochberg.

Pour résumer, en considérant que l'ajout des données mRNA-seq aux données cliniques classiques est pertinentes pour la prédiction si et seulement si au moins une des métriques d'évaluation est améliorée de manière significative, les données d'expressions génétique sont intéressantes pour prédire la survie pour 14 cancers sur 26 (ACC, BLCA, BRCA, CESC, HNSC, KIRC, KIRP, LGG, LIHC, MESO, PAAD, PRAD, THYM, UVM). Cependant, ces résultats sont à nuancer car nous avons utilisé uniquement des variables cliniques classiques, et aucune variable spécifique des cancers étudiés (*e.g.* œstrogènes pour BRCA, VOLKMANN et collab. [2019], HERRMANN et collab. [2020]). De plus, au-delà de la simple prédiction, les données mRNA-seq peuvent être utilisées pour différentes tâches [LÓPEZ DE MATURANA et collab., 2016] :

— stratifier les patients suivant les profils transcriptomiques [RICKETTS et collab., 2018].

— identifier des marqueurs prédictifs de réponse aux traitements [TERNÈS et collab., 2017].

— identifier de potentielles cibles thérapeutiques [WEI et collab., 2017].

Enfin, il est intéressant de remarquer que pour MESO, bien que le stade (TNM), l'âge et le sexe soient disponibles, les prédictions obtenues à partir de ces données cliniques sont mauvaises (C-index médian de 0.51). En revanche, pour LAML et GBM, avec uniquement l'âge et le genre comme variables cliniques prédictives de la survie, nous observons des C-index médians de 0,66 et de 0,60 respectivement. Ces résultats sont concordants avec

ceux obtenus par MILANEZ-ALMEIDA et collab. [2020] pour MESO et LAML, et l'étude des données cliniques n'a pas été mené pour GBM dans leur papier.

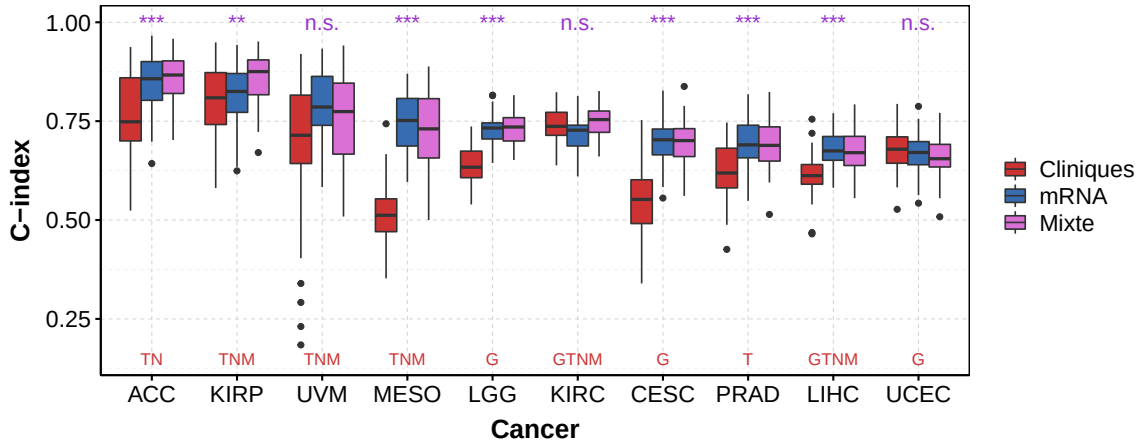


FIGURE 2.5 – **C-index obtenus avec les données cliniques, mRNA-seq, et en mixant les deux types de données.**

Pour chaque cancer, 50 C-index sont calculés en utilisant jusqu'à 6 variables cliniques (âge, genre, grade, T, N, M - rouge), les données mRNA-seq (bleu), et en mixant les deux types de données (*i.e.* les 6 variables cliniques et l'indice pronostique calculé avec les données mRNA-seq sont utilisés - violet). Pour chaque cas, les 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5).

L'âge est disponible pour tous les cancers et le genre pour 22 cancers sur 26 (*i.e.* non disponible pour CESC, PRAD, TGCT et UCEC). Les lettres rouges en bas du graphique indiquent la présence ou non des autres variables cliniques (G : grade, T : « *Tumor* », N : « *Node* », M : « *Metastasis* »).

Les étoiles en haut du graphique représentent les niveaux de significativité (légende ci-dessous) de tests de Wilcoxon unilatéraux corrigés par la méthode de Benjamini-Hochberg suivant l'ensemble des 26 cancers. Ces tests permettent d'observer si le C-index médian obtenu avec le mixte des données cliniques et mRNA-seq (violet) est significativement supérieur à celui obtenu avec uniquement les données cliniques (rouge)

n.s : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

2.4 Procédure de simulation des données de survie et d'évaluation des pénalisations

2.4.1 Procédure de simulation

Nous avons construit une procédure de simulation afin d'obtenir des données aussi proches que possible des données réelles. L'idée est d'évaluer les différentes pénalisations du modèle de Cox dans des conditions contrôlées et favorables (*i.e.* simulées avec le même modèle de risques proportionnels).

Tout d'abord, la structure de corrélation des données mRNA-seq est complexe et propre à chaque cancer [ARAN et collab., 2015]. Ainsi, nous avons utilisé les données mRNA-seq réelles de TCGA, et ces données ne sont donc pas simulées. Pour réduire la dimension, nous avons fait une première étape de pré-filtrage afin de ne garder que les 1 000 gènes qui ont les plus faibles p-valeurs du modèle de Cox univarié. Cette approche intuitive et simple permet de réduire le nombre de prédicteurs aux 1 000 gènes les plus corrélés à la survie, et demeure souvent utilisée en pratique [JIANG et collab., 2016; ZHAO et collab., 2015]. Nous avons utilisé les niveaux réels standardisés d'expression de ces 1 000 gènes pour simuler les temps de survie, et nous avons appliqué les méthodes de pénalisation (*i.e.* ridge, lasso, elastic net, adaptive elastic net) sur ces données transcriptomiques réelles et les temps de survie simulés.

Nous avons simulé les données de survie en différentes étapes décrites ci-dessous :

1. Choix au hasard des N gènes qui auront des coefficients β_j du modèle de Cox non nuls. Ces gènes sont ainsi corrélés à la survie dans les simulations et constituent la « vérité terrain ». Le nombre N de gènes avec des coefficients β_j non nuls s'échelonnera suivant les valeurs 10, 25, 50, 100 et 1 000.

2. Définition de la fonction de risque de base h_0. Pour cela un modèle de Cox-Weibull [KALBFLEISCH et PRENTICE, 2011] est appris sur les données réelles. Dans ce modèle, on suppose que le temps de survie de base T_0 suit une loi de Weibull de paramètres r et s. Sous cette hypothèse, la fonction de risque de base $h_0 = f_0/S_0$ (partie 1.6.4) est de la forme $h_0(t) = rst^{s-1}$. Nous avons ainsi estimé les deux paramètres de la loi de Weibull sur les données réelles de TCGA, et nous les avons utilisé pour les simulations.

3. Choix des coefficients β_j pour les gènes de la vérité terrain, tirés aléatoirement suivant une loi $\mathcal{N}(0, \sigma^2)$, avec $\sigma = 0,1\,;\, 0,25\,;\, 0,5\,;\, 1$.

   Ces deux premières étapes permettent de choisir le vecteur de coefficients $\boldsymbol{\beta}$, et nous pouvons définir un indice pronostique « oracle » comme la somme des $\beta_j X_j, i = j, \ldots, 1\,000$, pour chaque patient du jeu de données test. La notion d' « Oracle » renvoie au cas où le vecteur des coefficients $\boldsymbol{\beta}$ est connu, et nous l'utiliserons dans la suite pour définir le « C-index oracle », la « p-valeur oracle », et « l'IBS oracle » (partie 2.4.1.

4. Simulation des temps de survie T_i pour chaque patient $i = 1, \ldots, n$ [BENDER et collab., 2005]. En notant $F^{(i)}$ la fonction de répartition de la variable aléatoire du temps de survie $T^{(i)}$ pour le patient i, on a : $F^{(i)}(t|\mathbf{X}^{(i)}) = 1 - \exp(-H_0(t)\exp(\boldsymbol{\beta}^T\mathbf{X}^{(i)}))$. Ensuite, la variable aléatoire $1 - U = F^{(i)}(T^{(i)}|\mathbf{X}^{(i)})$ est uniforme et prend ses valeurs entre 0 et 1 [MOOD, 1974]. Finalement, en combinant les deux équations ci-dessus, on obtient :

   $T^{(i)} = H_0^{-1}(-\log(U)\exp(-\boldsymbol{\beta}^T\mathbf{X}^{(i)}))$.

En supposant que le temps de survie de base T_0 suit une loi de Weibull de paramètres r et s, la fonction de hasard cumulé de base est de la forme $H_0(t) = r\,t^s$, et on a :

$$T^{(i)} = \left(-\frac{\log(U)\exp(-\boldsymbol{\beta}^T \mathbf{X}^{(i)})}{r} \right)^{1/s}$$

5. Simulations des temps de censure C_i pour chaque patient $i = 1, \ldots, n$ [WAN, 2017]. Les temps de censure sont tirés suivant une loi uniforme de paramètres 0 et θ. Nous estimons le paramètre θ afin d'obtenir le même taux de censure que le jeu de données réel.

Ainsi, à la fin de ce processus de simulation, le temps de suivi t_i est défini par $t_i = \min(T_i, C_i)$, avec T_i le temps de survie et C_i le temps de censure, et le statut par $\delta_i = 1_{T_i \leq C_i}$, pour $i = 1, \ldots, n$.

2.4.2 Exemple de données de survie simulées

Les temps de survie obtenus sont représentés dans la Figure 2.6 pour LGG. Nous avons séparé les patients en quatre groupes de tailles égales suivant le quartile inférieur, la médiane, et le quartile supérieur des indices pronostiques oracles. Plus les courbes sont distinctes, et plus la séparation des patients en quatre groupes est pertinente en terme de risque associé à la survie.

Lorsque $\sigma = 0,1$ et $N = 10$, les gènes sont peu corrélés à la survie, et les indices pronostiques ne permettent pas de bien distinguer les patients en terme de risque (Fig. 2.6.C). Au contraire, lorsque l'on augmente σ ($\sigma = 1$) sans toucher à N ($N = 10$), l'intensité moyenne des coefficients $|\beta_j|$ est plus importante et la corrélation des indices pronostiques avec la survie est plus grande. La séparation entre les patients est alors plus efficace, et les courbes de survie sont plus distinctes (Fig. 2.6.D).

De la même manière, lorsque l'on augmente N ($N = 1\,000$) sans toucher à σ ($\sigma = 0.1$), un nombre plus importants de gènes est corrélé à la survie, et les courbes de survie sont plus distinctes (Fig. 2.6.A). Enfin, dans le cas extrême ou $N = 1\,000$ et $\sigma = 1$, l'ensemble des 1 000 gènes sont fortement corrélés à la survie, et on distingue une forte séparation entre les patients (*i.e.* presque aucun décès pour les patients qui ont un indice pronostique supérieur à la médiane, et décès dès la première année pour les patients qui ont un indice pronostique inférieur au premier quartile) (Fig. 2.6.B).

Cela peut s'expliquer de la manière suivante. Les indices pronostiques oracles (partie 2.4.1) sont de la forme $PI^{(i)}_{ora} = \sum_{j=1}^{N} \beta_{(j)} X^{(i)}_{(j)}$, avec $i = 1, \ldots, n$ l'indice d'un patient, $\beta_{(1)}, \ldots, \beta_{(N)}$ les N coefficients tirés suivant une loi $\mathcal{N}(0, \sigma^2)$, et $X^{(i)}_{(j)}$ le niveau d'expression standardisé du gène (j) pour le patient i. Les $X_{(j)}$ sont déterministes et les $\beta_{(j)}$ sont tirés de manière indépendantes, donc la variance des PI oracles est de la forme :

$$Var(\mathrm{PI}_{ora}^{(i)}) = \sigma^2 \sum_{j=1}^{N} X_{(j)}^2$$

La variance des indices pronostiques peut être vue comme l'intensité de la corrélation entre données d'expression génétique et survie, ou encore comme la quantité d'information contenue dans les données d'expression génétique relative à la survie. Cette quantité est de l'ordre de $N\sigma^2$, à une constante près. Lorsque $\sigma = 0.1$ et $N = 1\,000$, cette quantité d'information est de l'ordre de 10 (A) ; lorsque $\sigma = 1$ et $N = 1\,000$, cette quantité d'information est de l'ordre de 1 000 (B) ; lorsque $\sigma = 0.1$ et $N = 10$, cette quantité d'information est de l'ordre de 0,001 (C) ; lorsque $\sigma = 1$ et $N = 10$, cette quantité d'information est de l'ordre de 10 (D) (Fig. 2.7). Ces différents ordres de grandeur dans la quantité d'information expliquent donc les séparations plus ou moins distinctes des courbes de survie que nous observons sur la Figure 2.6. En particulier, les courbes de survie des cas $\sigma = 0,1$, $N = 1\,000$ et $\sigma = 1$, $N = 10$ contiennent la même quantité d'information et sont très similaires (Fig. 2.6 A et D, respectivement).

De la même manière, nous avons calculé les courbes de Kaplan-Meier sur données réelles avec les indices pronostiques estimés avec un modèle de Cox avec pénalisation elastic net (2.6.E). Nous voyons que la séparation des patients est intermédiaire entre les courbes de la Figure 2.6.B et 2.6.C. Nous avons en effet choisi les paramètres N et σ pour obtenir des simulations de temps de survie dont les résultats englobent les cas réels, et pour que les cas $\sigma = 0,1$, $N = 1\,000$ (σ minimum et N maximum) et $\sigma = 1$, $N = 10$ (σ maximum et N minimum) contiennent la même quantité d'information. Enfin, avec les paramètres $\sigma = 0,25$, $N = 25$, les PI réels et simulés sont comparables (Fig. 2.7.F) et, si l'on s'arrête à l'horizon 5 ans, les courbes de survie le sont aussi (Fig. 2.6.F).

2.4.3 Évaluation des méthodes de pénalisation sur données simulées

Pour affiner et compléter les résultats obtenus sur données réelles, nous avons étudié les capacités de prédiction des différentes pénalisations du modèle de Cox (*i.e.* ridge, lasso, elastic net et adaptive elastic net) sur données simulées. Les simulations nous permettent aussi d'étudier les capacités de sélection des pénalisation lasso, elastic net et adaptive elastic net. Nous avons résumé le processus d'évaluation sur la Figure 2.8.

Concernant les prédictions, pour une simulation donnée, nous pouvons calculer un « C-index oracle » ($\hat{C}_{ora}$), une « p-valeur du modèle de Cox oracle » ($\hat{p}_{ora}$), et un « IBS oracle » ($\hat{\mathrm{IBS}}_{ora}$) à partir des « indices pronostiques oracles » du jeu de données test simulé (partie 2.4.1 et Fig. 2.8). Ces indicateurs peuvent être considérés comme les métriques de prédiction « optimales » que l'on peut atteindre avec le modèle. Nous pouvons ainsi établir une comparaison de ces métriques oracles avec les métriques obtenues après estimation du vecteur des coefficients $\boldsymbol{\beta}$ par le modèle de Cox pénalisé. Par exemple, pour le C-index, nous calculons la différence $\hat{C}_{ora} - \hat{C}$ entre le C-index oracle et le C-index obtenu après

estimation des coefficients β_j par le modèle de Cox pénalisé, pour chaque répétition (Fig. 2.8). Cette procédure permet donc d'évaluer dans quelle mesure les prédictions obtenues si les coefficients β_j étaient connus sont atteintes.

Concernant les performances de sélection, nous avons choisi la proportion de faux positifs (*False Discovery Proportion*, FDP), qui se définit comme la proportion de faux positifs parmis les gènes sélectionnés, et la sensibilité, qui se définit comme la proportion de gènes de la vérité terrain qui sont sélectionnés :

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

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

avec « FP » le nombre de faux positifs, « VP » le nombre de vrais positifs, et « FN » le nombre de faux négatifs. Nous définissons la vérité terrain comme les N gènes tirés aléatoirement dans le processus de simulation, et qui ont servi à simuler les temps de survie (partie 2.4.1).

Dans les prochaines parties de ce chapitre, nous nous attacherons à étudier les capacités de prédiction et de sélection du modèle de Cox pénalisé sur données simulées.

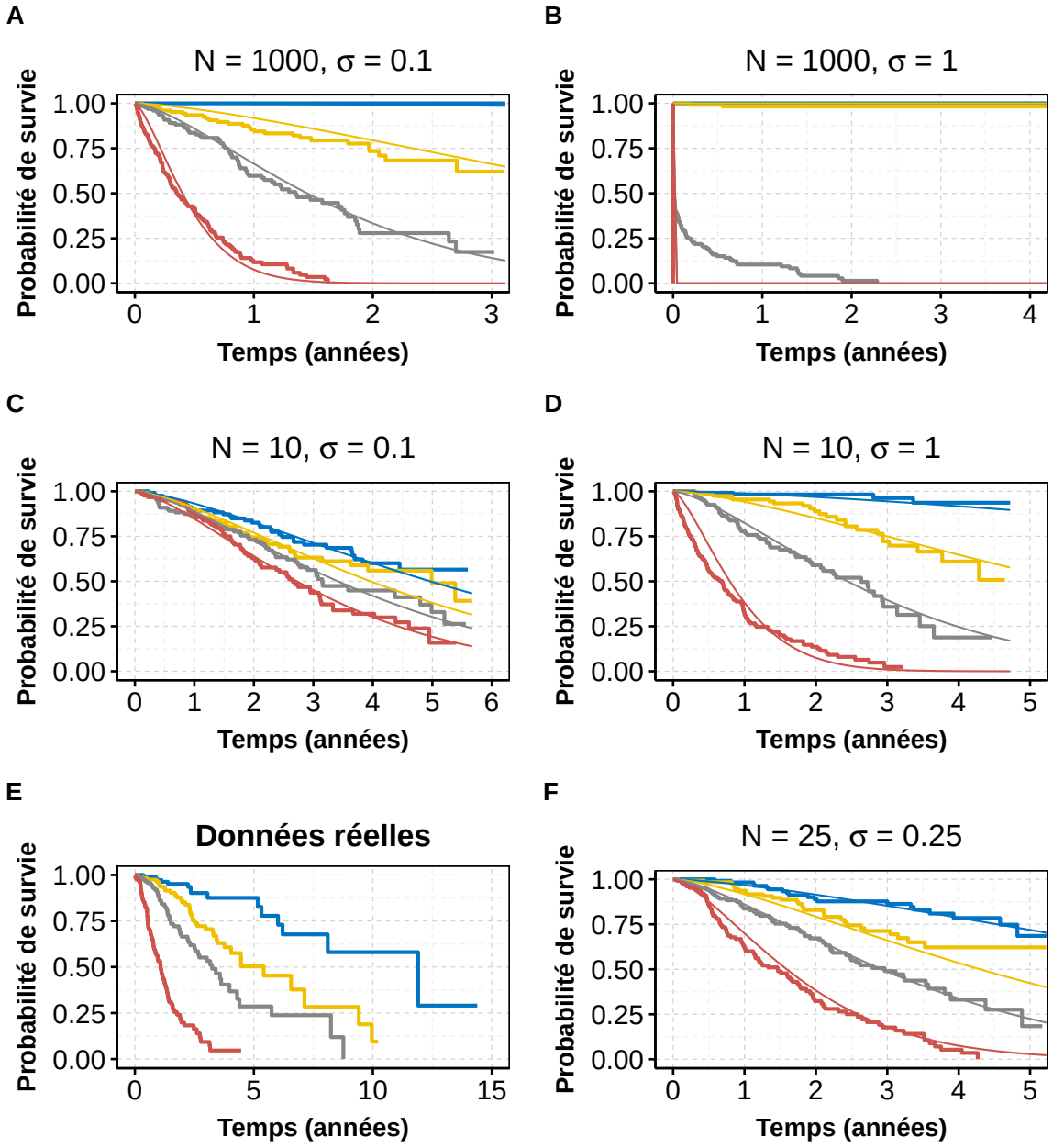


FIGURE 2.6 – **Courbes de Kaplan-Meier des données de survie obtenues avec différents paramètres de simulation et sur données réelles pour LGG.**

Les données de survie sont simulées avec N = 1 000 et σ = 0.1 (A), N = 1 000 et σ = 1 (B), N = 10 et σ = 0.1 (C), N = 10 et σ = 1 (D), N = 25 et σ = 0.25 (F), suivant la procédure décrite à la partie 2.4.1. Les courbes de Kaplan-Meier obtenues après estimation des coefficients β_j par elastic net sur données réelles sont aussi tracées (F).

Dans chaque graphique, les quatre courbes correspondent aux courbes de survie obtenues en séparant les patients suivant leurs indices pronostiques « oracles » pour les simulations (A, B, C, D, F), et estimés par le modèle de Cox avec pénalisation elastic net pour les données réelles (E). Pour les graphiques A, B, C, D et F, nous avons ajouté les courbes de survie du modèle de Cox-Weibull calculées avec les PI médians de chaque groupe (courbes lisses).

Bleu : PI inférieurs au premier quartile, jaune : PI compris entre le premier quartile et la médiane, gris : PI compris entre la médiane et le troisième quartile, rouge : PI supérieurs au troisième quartile.

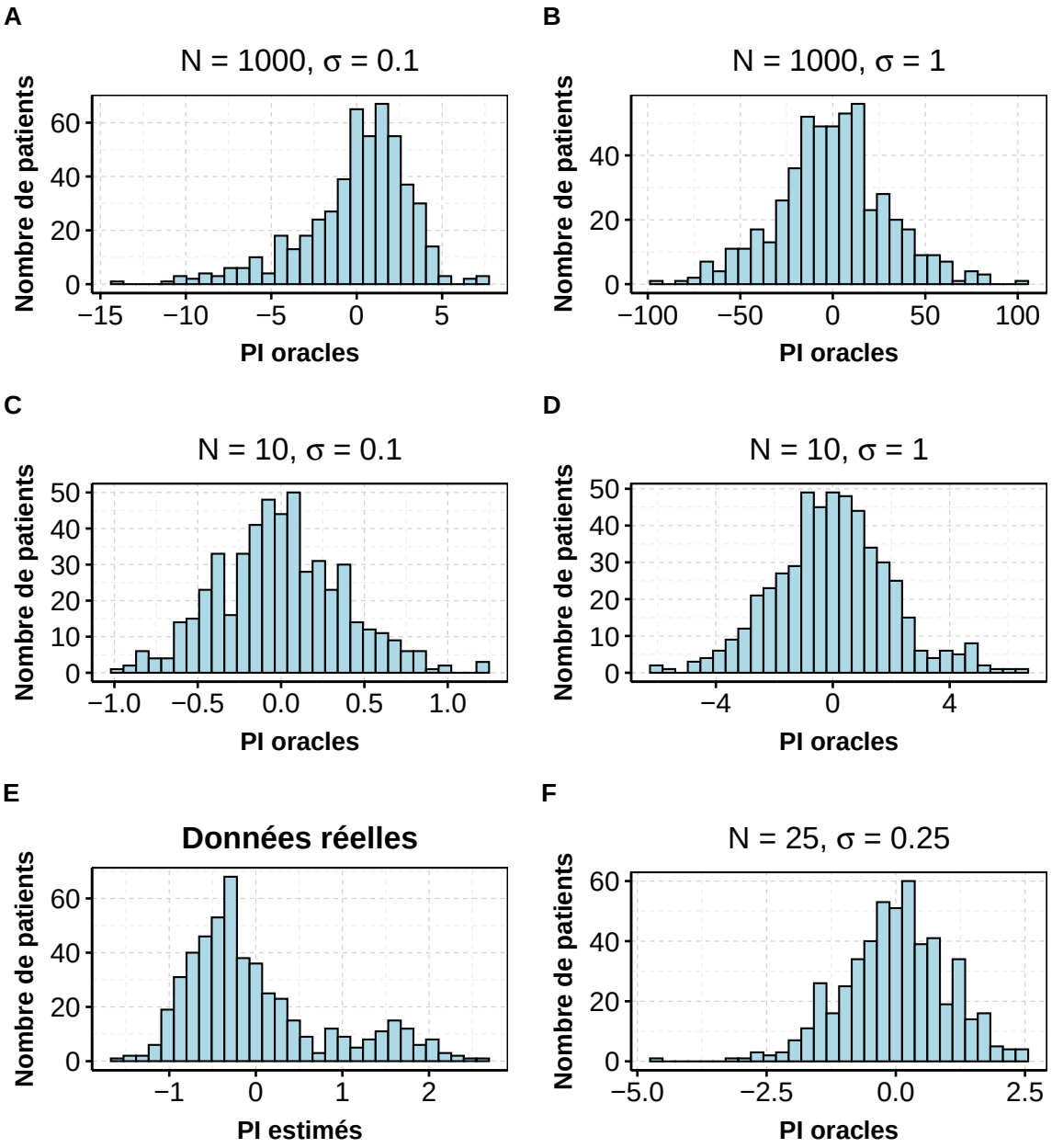


FIGURE 2.7 – **Histogrammes des indices pronostiques oracles obtenus avec différents paramètres de simulation et sur données réelles pour LGG.**

Les données de survie sont simulées avec N = 1 000 et σ = 0.1 (A), N = 1 000 et σ = 1 (B), N = 10 et σ = 0.1 (C), N = 10 et σ = 1 (D), N = 25 et σ = 0.25 (F), suivant la procédure décrite à la partie 2.4.1. Les indices pronostiques obtenus après estimation des coefficients β_j par elastic net sur données réelles sont aussi tracées (F).

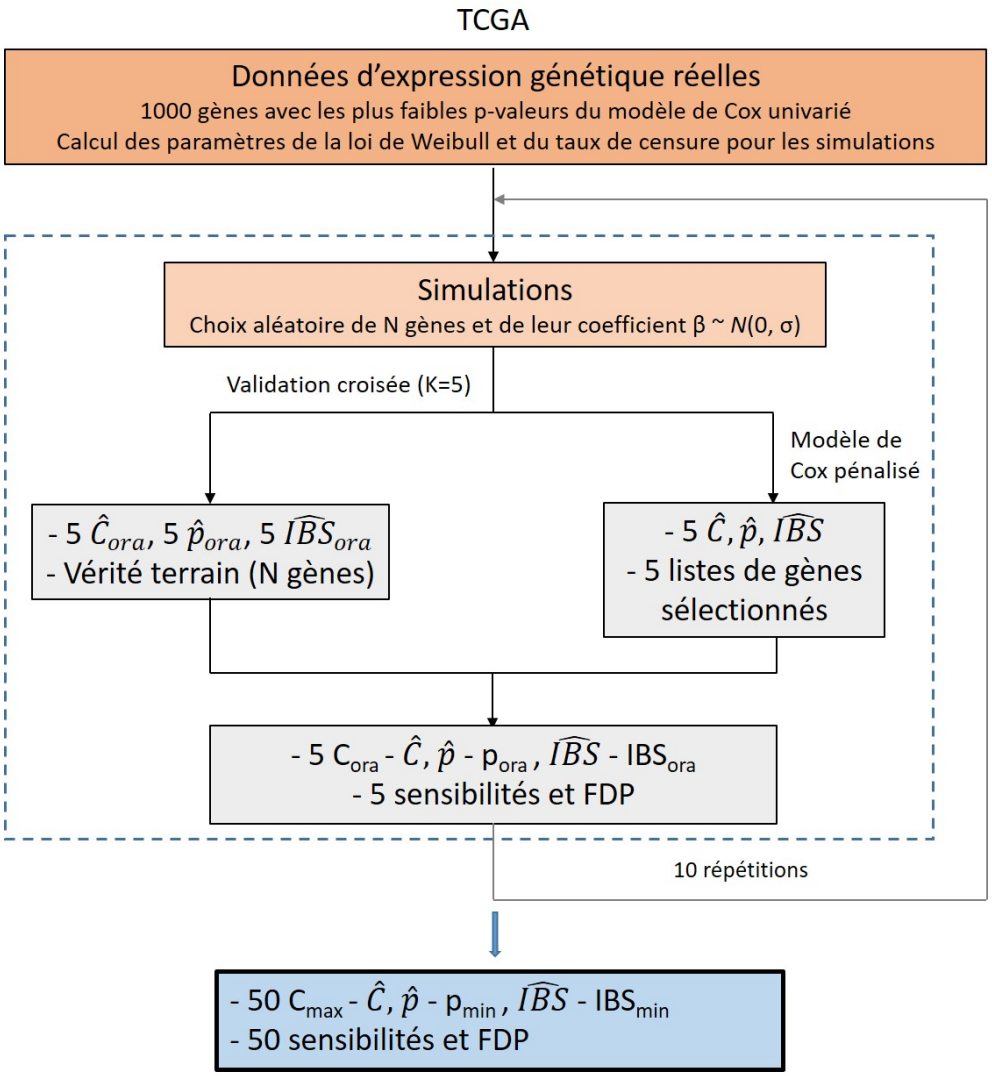


FIGURE 2.8 – **Schéma récapitulatif des procédures de simulation et d'évaluation du modèle de Cox pénalisé sur données simulées décrite dans la partie 2.4.1.**

2.5 Capacités de prédiction sur données simulées

2.5.1 Influence des paramètres de simulation sur les prédictions oracles

Tout d'abord, et de façon attendue, les prédictions oracles sont croissantes en fonction de l'écart-type σ des coefficients oracles β_j non nuls, et du nombre N de gènes dans la vérité terrain (Fig. 2.9.A, et données non montrées pour les autres cancers et les autres métriques). Plus l'écart-type des coefficients β_j non nuls est important, et donc plus l'intensité de la corrélation entre niveaux d'expression et survie est élevée, plus les prédictions oracles sont bonnes. De plus, même dans le cadre favorable où les données de survie sont simulées suivant un modèle de Cox-Weibull, si l'intensité des coefficients β_j non nuls et le nombre de gènes N dans la vérité terrain ne sont pas suffisants, les prédictions oracles sont mauvaises (C-index < 0.6 pour l'ensemble des cancers, Fig. 2.9.A).

Notons que les C-index et les IBS oracles (Tab. 2.3) et estimés par le modèle de Cox pénalisé sont très similaires suivant les jeux de données. Les C-index et les IBS oracles se répartissent autour de 0,89 et 0,09 respectivement. En revanche, les p-valeurs du modèle de Cox univarié diffèrent grandement d'un cancer à un autre. Ces différences s'expliquent en partie par l'influence du nombre d'échantillons sur la p-valeur (Fig. 2.10). La corrélation de Spearman entre le nombre de patients dans l'échantillon et la médiane des p-valeurs du modèle de Cox univarié (-log10) oracles est de 0.90 (p-valeur < 0.001, test de corrélation de Spearman). Ainsi, la p-valeur du modèle de Cox univarié permet de comparer des algorithmes sur un même jeu de données ou sur des jeux de données comportant un nombre d'échantillons similaires, mais pas à comparer des résultats obtenus sur des cohortes de tailles différentes. Ces corrélations ne sont pas observées pour le C-index et l'IBS (0,43 pour le C-index, et -0,25 pour l'IBS).

TABLEAU 2.3 – **Médianes des C-index oracles, des p-valeurs du modèle de Cox univarié (-log10) oracles, et des IBS oracles pour les 26 cancers de TCGA étudiés - données simulées.**
Les médianes sont calculées sur l'ensemble des simulations (σ = 0,1; 0,25; 0,5; 1 et N = 10; 25; 50; 100; 1 000).

Cancer	ACC	BLCA	BRCA	CESC	COAD	ESCA	GBM	HNSC	KIRC	KIRP	LAML	LGG	LIHC
C-index	0,83	0,89	0,92	0,91	0,91	0,9	0,86	0,89	0,89	0,91	0,87	0,88	0,9
-log10(p)	2,9	18	21	8,7	14	7,2	9,8	22	20	5,6	8,3	19	14
IBS	0,12	0,092	0,06	0,082	0,081	0,088	0,067	0,098	0,096	0,076	0,091	0,1	0,093
Cancer	LUAD	LUSC	MESO	OV	PAAD	PRAD	READ	STAD	TGCT	THCA	THYM	UCEC	UVM
C-index	0,9	0,89	0,85	0,88	0,87	0,9	0,9	0,89	0,89	0,88	0,89	0,91	0,83
-log10(p)	20	23	5,4	17	8,3	19	5	16	4,9	20	4	18	2,6
IBS	0,093	0,095	0,079	0,079	0,11	0,085	0,092	0,1	0,1	0,087	0,098	0,077	0,11

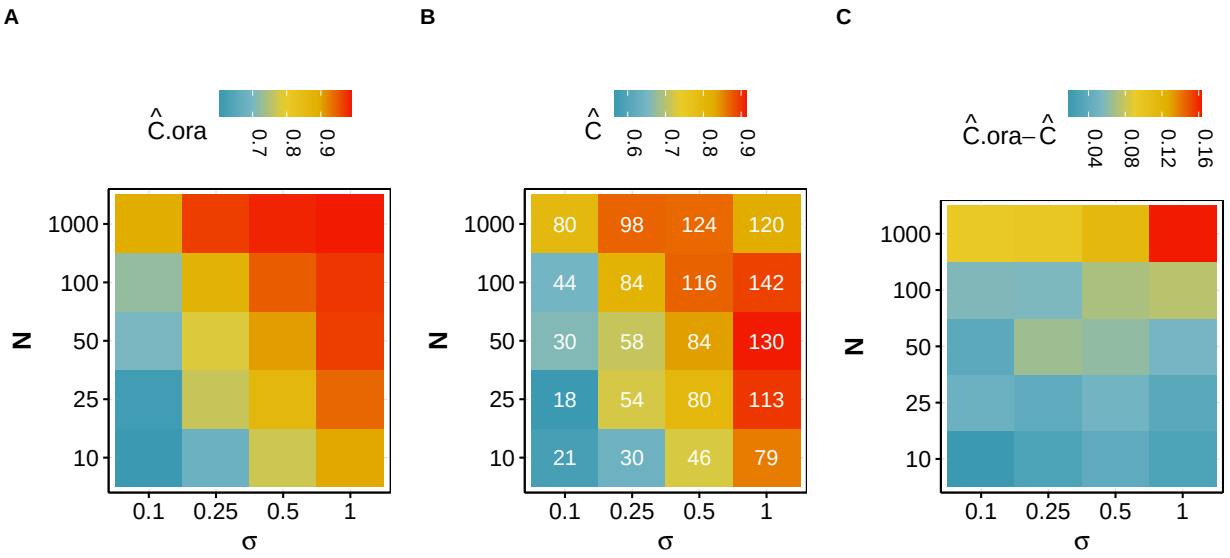


FIGURE 2.9 – **C-index médians oracles (A) et estimés (B), et différences médianes entre les C-index oracles et estimés (C) pour différents paramètres de simulation pour LGG et la pénalisation elastic net - données simulées.**

Les C-index sont calculés par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$).

$\hat{C}$: C-index estimé par le modèle de Cox pénalisé ; $\hat{C}$.ora : C-index calculé avec les indices pronostiques oracles utilisés pour les simulations. Le nombre médian de gènes sélectionnés par la pénalisation elastic net est indiqué en blanc pour chaque paramètre de simulation dans le graphique B.

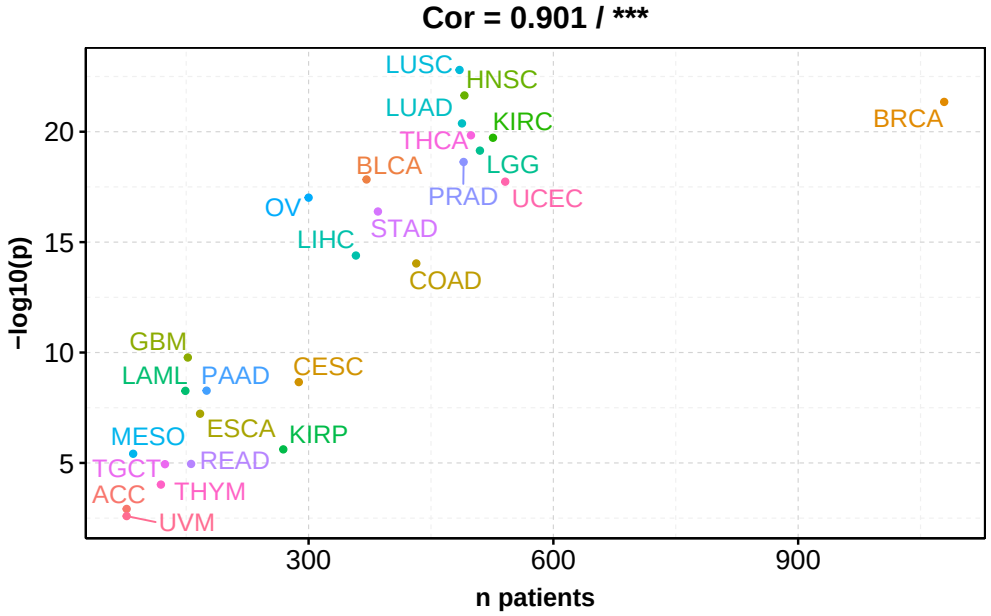


FIGURE 2.10 – **Médiane des p-valeurs du modèle de Cox univarié (-log10) oracles en fonction du nombre de patients pour les 26 cancers de TCGA étudiés - données simulées.**

Les p-valeurs oracles du modèle de Cox univarié sont calculées par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$), et la médiane est calculée.

2.5.2 Influence des paramètres de simulation sur les prédictions obtenues avec le modèle de Cox pénalisé

Tout d'abord, nous observons que de meilleures prédictions sont obtenues avec le modèle de Cox pénalisé lorsque σ et N augmentent, et cela pour chacune des trois métriques (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS) (Fig. 2.9.B, et données non montrées pour les autres cancers, les autres pénalisations, et les autres métriques). De plus, plus sigma augmente, plus le nombre de gènes sélectionnés augmente et tend à être sur-estimé (sauf pour le cas N = 1 000).

Ensuite, pour tous les jeux de données et les trois métriques d'évaluation des prédictions (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS), nous observons qu'il est préférable d'avoir peu de gènes fortement reliés à la survie (σ = 1 et N = 10, case en bas à droite sur les graphiques de la Figure 2.9), que de nombreux gènes marginalement associés à la survie (σ = 0.1 et N = 1 000, case en haut à gauche sur les graphiques de la Figure 2.9) pour les pénalisations lasso, elastic net et adaptive elastic net. En effet, ces deux scénarios de simulations contiennent la même quantité d'information (10), et les C-index médians obtenus avec peu de gènes fortement corrélés à la survie (σ = 1 et N = 10) sont plus élevés pour tous les cancers sauf pour KIRP (Fig. 2.11.A). Nous observons des résultats similaires pour l'IBS et la p-valeur du modèle de Cox (Fig. Annexe A.6). Ces résultats ne sont en revanche pas observés pour la pénalisation ridge : la qualité des prédictions obtenues parait comparable entre ces deux cas « extrêmes » (Fig. 2.11.B). La pénalisation ridge semble moins bien adaptée lorsque peu de gènes sont fortement reliés à la survie (σ = 1 et N = 10), que les pénalisations permettant de sélectionner un sous-ensemble de gènes (*i.e.* lasso, elastic net, adaptive elastic net).

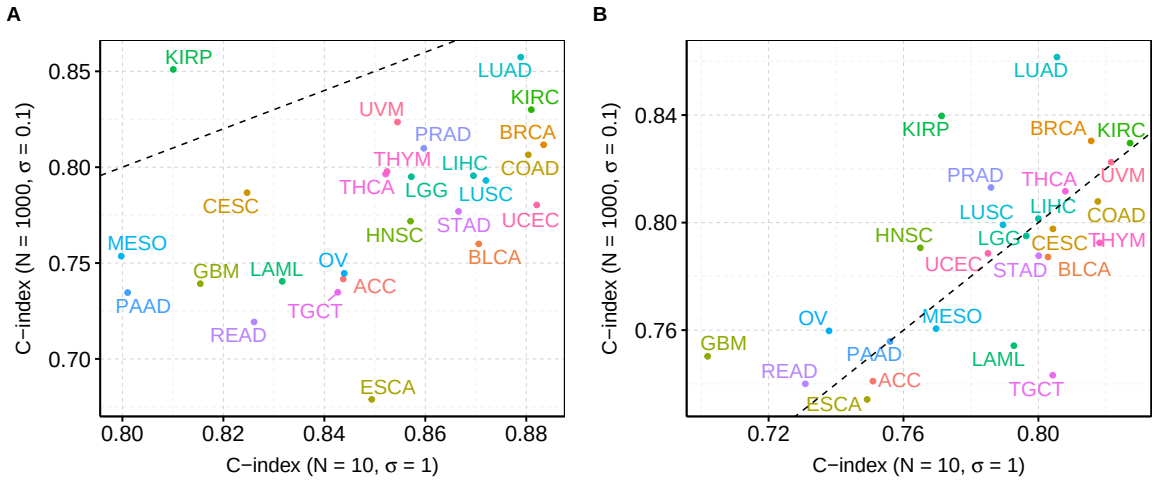


FIGURE 2.11 – **C-index médians obtenus avec elastic net (A) et ridge (B) pour différents paramètres de simulations contenant la même quantité d'information pour les 26 cancers de TCGA étudiés - données simulées.**
Les C-index médians sont obtenus avec N = 1 000 et σ = 0.1 en ordonnée, l, et avec N = 10 et σ = 1 en abscisse.

2.5.3 Comparaison des prédictions oracles et des prédictions obtenues par le modèle de Cox pénalisé

Tout d'abord, nous observons que les métriques d'évaluation des prédictions (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS) obtenues avec le modèle de Cox pénalisé sont, à quelques exceptions près, moins bonnes que les métriques oracles (Fig. 2.12, données non montrées pour les autres cancers et les autres pénalisations). Ceci souligne que les prédictions obtenues avec le modèle de Cox pénalisé sont « conservatives » : on estime un effet moins fort (C-index plus faible, ou p-valeur et IBS plus élevés) que celui qui existe réellement, ou qui est estimé ici par l'Oracle. Néanmoins, ce biais qui consiste à sous-estimer les effets réels, est plutôt dans le bon sens pour interpréter nos découvertes. Si l'on observe un C-index élevé (resp. une p-valeur du modèle de Cox faible, un IBS faible) avec notre algorithme sur jeu de données réelles, on peut être confiant sur l'information et les capacités prédictives des données.

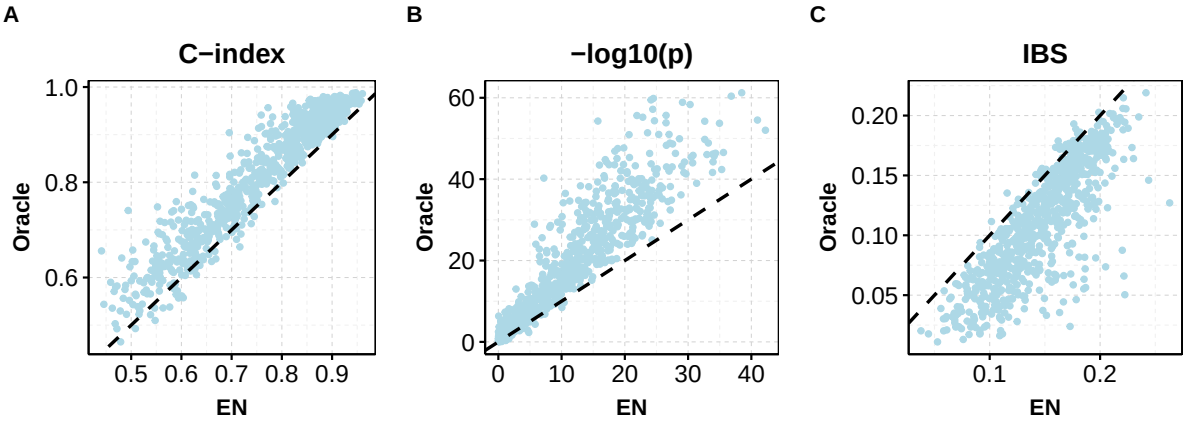


FIGURE 2.12 – **Métrique oracle en fonction de la métrique estimée par le modèle de Cox avec pénalisation elastic net pour LGG - données simulées.**
(A) C-index, (B) p-valeur du modèle de Cox univarié (-log10), (C) IBS.
Chaque point correspond à une estimation de la métrique considérée pour des paramètres de simulation données ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$). Pour chaque couple de paramètres (σ, N), les métriques d'évaluation des prédictions sont calculées par 10 répétitions d'une validation croisée (K=5).
EN : elastic net.

Ensuite, notons que les performances de prédiction obtenues sur données simulées dépendent des jeux de données (Tab. 2.4 et Tab. Annexe A.1 et A.2). Par exemple, la différence médiane entre C-index oracles et C-index obtenus par le modèle de Cox avec pénalisation lasso sur les mêmes données simulées est de 0,16 pour ESCA, et de 0,054 pour BRCA, soit une différence d'un facteur 3. Ces disparités peuvent venir des différents paramètres utilisés pour simuler les temps de survie (*i.e.* paramètres de la loi de Weibull, partie 2.4.1), de la structure de corrélation des jeux de données (*e.g.* existence de sous-types de cancers), ou encore du nombre de patients différents entre les jeux de données (partie 2.5.1, Fig. 2.13.A). En revanche, le taux de censure n'a que peu d'influence sur les

prédictions obtenues (Fig. 2.13.B). Les deux dernières conclusions se vérifient aussi pour la p-valeur du modèle de Cox et l'IBS (Fig. Annexe A.7 et A.8 respectivement).

TABLEAU 2.4 – **Médianes des différences entre C-index oracles et estimés par le modèle de Cox pénalisé pour les 26 cancers de TCGA étudiés - données simulées.**
Les médianes sont calculées sur l'ensemble des simulations ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$).
EN : elastic net ; AEN : adaptive elastic net.

Cancer	ridge	lasso	EN	AEN
ACC	0,11	0,14	0,12	0,13
BLCA	0,11	0,084	0,086	0,082
BRCA	0,1	0,054	0,057	0,059
CESC	0,12	0,12	0,12	0,11
COAD	0,12	0,094	0,097	0,092
ESCA	0,16	0,16	0,16	0,14
GBM	0,14	0,13	0,13	0,12
HNSC	0,11	0,069	0,073	0,07
KIRC	0,082	0,059	0,06	0,057
KIRP	0,11	0,12	0,11	0,11
LAML	0,11	0,12	0,11	0,11
LGG	0,083	0,057	0,06	0,058
LIHC	0,1	0,087	0,086	0,087
LUAD	0,1	0,068	0,072	0,071
LUSC	0,11	0,073	0,077	0,074
MESO	0,14	0,13	0,13	0,13
OV	0,14	0,11	0,11	0,1
PAAD	0,11	0,11	0,11	0,1
PRAD	0,094	0,072	0,074	0,072
READ	0,14	0,16	0,14	0,14
STAD	0,11	0,079	0,083	0,08
TGCT	0,13	0,13	0,13	0,12
THCA	0,095	0,07	0,071	0,07
THYM	0,12	0,13	0,11	0,11
UCEC	0,12	0,08	0,081	0,081
UVM	0,088	0,11	0,095	0,11

Les pénalisation lasso, elastic net et adaptive elastic net atteignent des performances de prédictions équivalentes (Tab. 2.4). En revanche, la pénalisation ridge obtient en moyenne des résultats clairement dégradés en terme de qualité de prédiction. Par exemple, les différences médianes entre C-index oracles et estimés sont au moins 1,5 fois supérieurs à au moins une des trois autres pénalisations (*i.e.* ridge, lasso, elastic net, adaptive elastic net) pour BRCA, HNSC et LUSC. La pénalisation ridge assigne un coefficient β non nul à l'ensemble des gènes, et lorsque que seul un sous-ensemble restreint de gènes est corrélé à la survie, les prédictions obtenues sont parfois moins bonnes que les autres méthodes de pénalisation qui permettent de sélectionner un sous-ensemble de gènes (*i.e.* lasso, elastic net, adaptive elastic net).

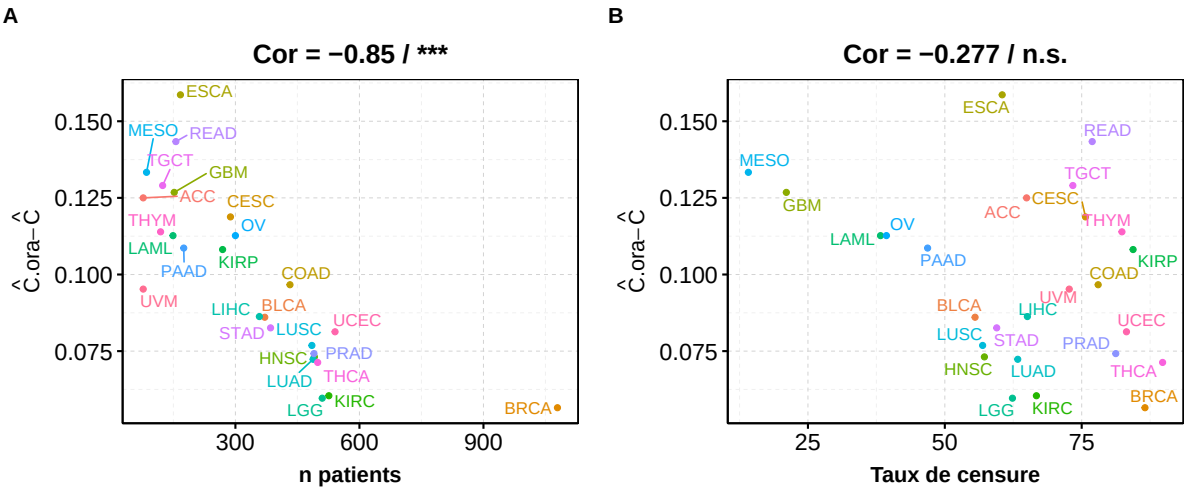


FIGURE 2.13 – **Différences médianes entre les C-index oracles et estimés en fonction du nombre de patients (A) et du taux de censure (B) pour l'ensemble des 26 cancers étudiés.**
Les corrélations et les niveaux de significativité d'un test de Spearman sont indiqués en haut des graphiques. (*** : p-valeur < 0,001 ; n.s. : non significatif). Les différences médianes sont calculées sur l'ensemble des simulations ($\sigma = 0,1;0,25;0,5;1$ et $N = 10;25;50;100;1\,000$).

Enfin, les capacités de prédictions oracles sont loin d'être atteintes pour certains cancers. Par exemple, pour ESCA et la pénalisation elastic net, la différence médiane entre C-index oracles et C-index obtenus par le modèle de Cox est de 0,16. De telles différences supérieures à 0,10 pour les quatre pénalisations étudiées (*i.e.* ridge, lasso, elastic net, adaptive elastic net) se retrouvent pour d'autres jeux de données (*e.g.* ACC, CESC, GBM, KIRP, LAML, MESO, OV, PAAD, READ, TGCT, THYM). En moyenne et pour la pénalisation elastic net, les différences médianes entre les C-index oracles et estimés sont de 0.099.

Ces résultats sous-optimaux de prédiction se retrouvent avec la p-valeur du modèle de Cox univarié (Tab. Annexe A.1) et l'IBS (Tab. Annexe A.2). Nous avons calculé le niveau de significativité de ces différences pour les trois métriques (*i.e.* C-index, p-valeur, IBS) par un test des rangs signés de Wilcoxon unilatéral, et les p-valeurs corrigées par Benjamini-Hochberg pour les 26 cancers sont toutes significatives au niveau 0,001, pour toutes les métriques et toutes les pénalisations.

L'existence de corrélations entre les gènes de la vérité terrain et d'autres gènes rend difficile la sélection des « vrais » prédicteurs. Dans les prochaines parties, nous allons évaluer les performances de sélection des pénalisations lasso, elastic net, et adaptive elastic, et nous étudierons l'influence des performances de sélection sur les capacités de prédiction.

2.6 Capacités de sélection des méthodes de pénalisation du modèle de Cox

2.6.1 Stabilité des gènes sélectionnés dans deux cohortes distinctes - données réelles

Dans cette sous-partie, nous allons nous intéresser à la stabilité des gènes sélectionnés (*i.e.* gènes avec des coefficients β non nuls dans le modèle de Cox pénalisé) sur deux jeux de données ne contenant aucun patient en commun. Pour cela, nous avons séparé les jeux de données de TCGA (données réelles) en deux sous-jeux de données contenant chacun 50% des patients, et nous avons appliqué un modèle de Cox pénalisé sur ces deux sous-jeux de données, indépendamment. Le pourcentage de gènes sélectionnés en commun dans les deux sous-jeux de données peut ainsi être calculé, et cette métrique permet d'estimer la stabilité de la sélection pour chacune des trois méthodes de pénalisation :

$$\text{Pourc}_{\text{GS}} = \frac{\#(\text{GS}_1 \cap \text{GS}_2)}{\min(\#\text{GS}_1, \#\text{GS}_2)},$$

avec Pourc_{GS} le pourcentage de gènes sélectionnés à la fois dans les jeux de données 1 et 2 ; GS_k l'ensemble des gènes sélectionnés dans le jeux de données k, $k = 1,2$; # la fonction cardinale.

Nous avons répété cette procédure 50 fois pour les trois pénalisations (*i.e.* lasso, elastic net, adaptive elastic net) et sur l'ensemble des cancers (Fig. 2.14 et Fig. Annexe A.9).

Tout d'abord, pour l'ensemble des méthodes et des jeux de données, le pourcentage médian de gènes sélectionnés en commun dans les deux sous-jeux de données comportant chacun 50% des patients n'excède pas 10% en moyenne (Fig. 2.14 et Fig. Annexe A.9). La sélection est donc très instable et les gènes sélectionnés sur un jeu de données dépendent des patients qui le compose.

Différentes hypothèses, biologiques et mathématiques, peuvent être émises quant à l'origine de cette instabilité [Domany, 2014; Ein-Dor et collab., 2005]. Comme nous l'avons vu en introduction, le « principe d'unicité du cancer » (partie 1.1.3), stipule que cette maladie est très hétérogène entre patients. Ainsi, afin d'obtenir des résultats plus robustes, la taille des cohortes de patients doit être suffisamment grande [Michiels et collab., 2005]. Ensuite, la présence de fortes corrélations dans les données d'expression génétiques peut induire une forte instabilité dans le sous-ensemble de gènes sélectionné. En effet, dans deux sous-jeux de données, des gènes différents mais fortement corrélés peuvent être sélectionnés.

Finalement, la stabilité des gènes sélectionnés est comparable pour elastic net et adaptive elastic net, et apparaît moins bonne pour lasso (Fig. 2.14 et Fig. Annexe A.9). En effet, la pénalisation lasso a tendance à ne sélectionner qu'une variable parmi un groupe de variables corrélées, alors que elastic net assigne des coefficients similaires, au signe près,

aux variables d'un tel groupe [Zou et Hastie, 2005].

Pour remédier à cela, la méthode « *stability selection* » [Meinshausen et Bühlmann, 2010] consiste à apprendre m modèles en faisant un tirage avec remise des patients (« *bootstrap* »), et de ne retenir que les gènes qui ont été sélectionné dans plus de x% des cas. Les paramètres m et x sont fixés par l'utilisateur suivant les objectifs de l'étude à mener. Cette méthode est ainsi intéressante à considérer pour obtenir plus de stabilité dans la sélection.

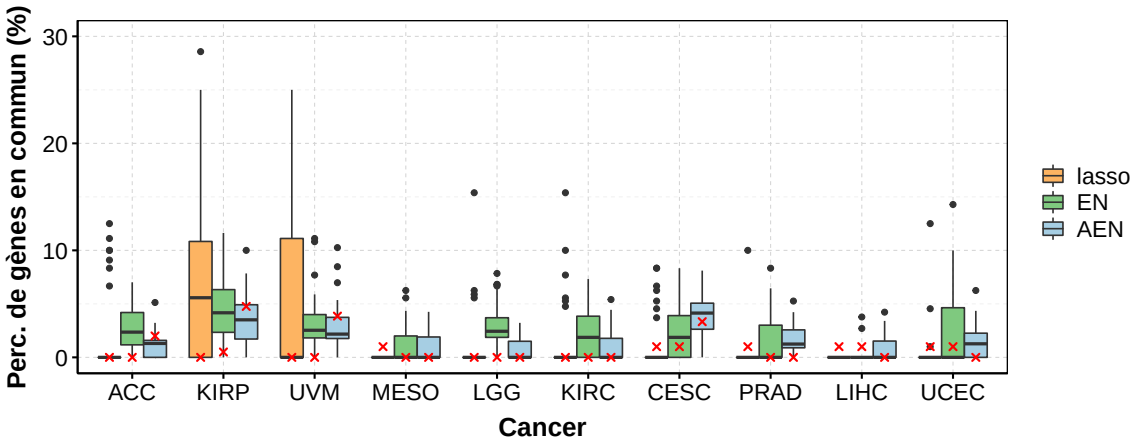


FIGURE 2.14 – **Stabilité des gènes sélectionnés.**
Pourcentage de gènes sélectionnés en commun dans deux sous-jeux de données comportant chacun 50% des patients (50 répétitions). Les *boxplots* sont les résultats obtenus avec λ_{min}, et les croix rouges sont les médianes des résultats obtenus avec λ_{1se}. Les 10 cancers qui ont le plus grand C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 16 autres cancers étudiés sont placés en annexe (Fig. Annexe A.9).

2.6.2 Impact des paramètres de simulations sur les performances de sélection

Dans cette sous-partie et dans la suite de cette partie, nous travaillerons sur des données simulées afin d'étudier la capacité des méthodes à recouvrir un sous-ensemble de gènes. Dans cette optique, nous excluons le cas N = 1 000 des scénarios de simulations, où l'ensemble des gènes du jeu de données constituent la vérité terrain. Ainsi, les paramètres que nous utilisons pour les simulations sont $\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100$.

Ainsi, la sensibilité est une fonction croissante de σ et décroissante de N, et le FDP est une fonction décroissante de σ et de N. La meilleure sensibilité et la plus faible proportion de fausse découverte (FDP) sont atteintes lorsque l'écart-type des coefficients β_j non nuls est grand (*i.e.* $\sigma = 1$), et lorsque la taille de la vérité terrain est petite pour la sensibilité (*i.e.* N = 10), et grande pour le FDP (*i.e.* N = 100) (Fig. 2.15). Cela peut s'expliquer par le fait que le nombre de gènes sélectionné augmente moins vite que le nombre de gènes présents dans la vérité terrain.

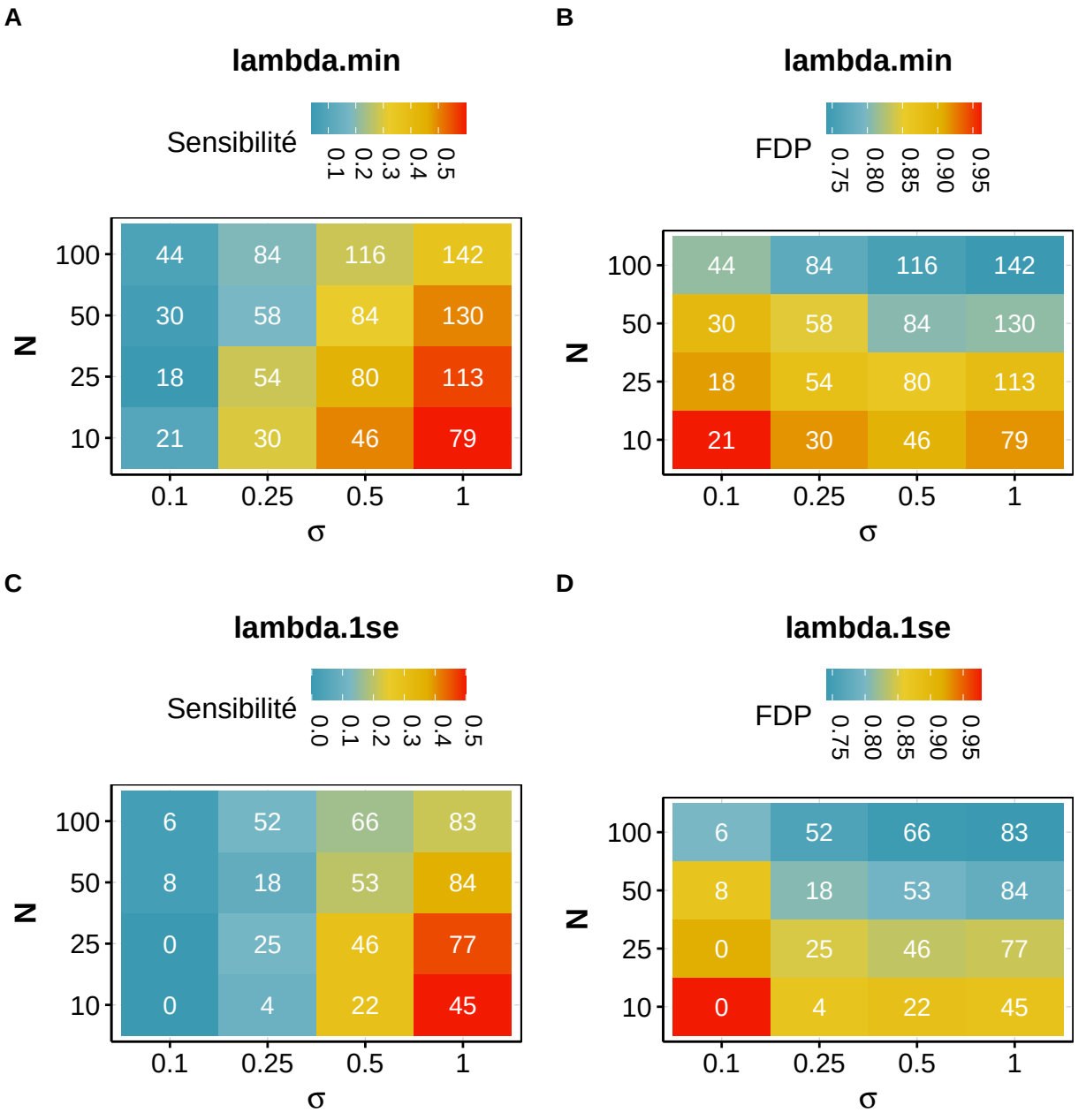


FIGURE 2.15 – **Sensibilités (A, C) et taux de fausses découvertes (FDP) (B, D) obtenus avec λ_{min} (A, B) et λ_{1se} (C, D) pour LGG et la pénalisation elastic net - données simulées.**
La sensibilité et la proportion de fausses découvertes (FDP) sont calculées par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100$). Le nombre médian de gènes sélectionnés est indiqué en blanc pour chaque cas.

2.6.3 Influence du poids accordé à la pénalisation

Nous avons décrit comment choisir le poids λ à assigner à la pénalisation dans la partie 1.8.3. La valeur de λ_{1se} est plus grande que celle de λ_{min}, donc le poids accordé à la pénalisation est plus important, et donc moins de gènes sont sélectionnés. Dans les simulations, nous observons que la sensibilité médiane et le FDP médian obtenus avec λ_{min} sont significativement supérieurs à ceux obtenus avec λ_{1se} pour elastic net (p-valeur <

0.01, test des rangs signés de Wilcoxon avec correction de Benjamini-Hochberg) (Fig. 2.16 A et B). Le nombre plus important de gènes retenus par λ_{min} permet de mieux recouvrir la vérité terrain (sensibilité plus élevé), mais accroît le taux de faux positifs. Les mêmes conclusions sont obtenues pour les pénalisations lasso et adaptive elastic net.

En revanche, le ratio sensibilité sur FDP est significativement plus avantageux pour λ_{min} avec la pénalisation elastic net (p-valeur < 0.001, test des rangs signés de Wilcoxon avec correction de Benjamini-Hochberg) (Fig. 2.16.C). Ainsi, le taux de faux positifs augmente moins vite que la sensibilité entre λ_{1se} et λ_{min} lorsque plus de gènes sont sélectionnés, et les mêmes conclusions sont obtenues pour les pénalisations lasso et adaptive elastic net. Formulé autrement, la baisse du FDP induite par le choix de λ_{1se} est contrebalancée par une baisse plus importante de la sensibilité.

De plus, dans la partie 2.2.1, nous avons montré que les performances de prédiction sont meilleures avec λ_{min} qu'avec λ_{1se}. Ainsi, dans la suite du manuscrit, nous utiliserons systématiquement λ_{min} comme poids assigné à la pénalisation.

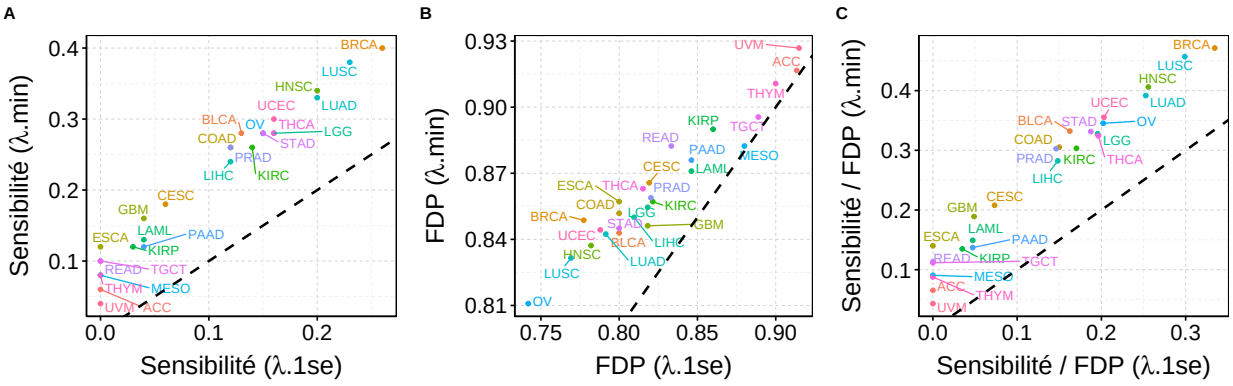


FIGURE 2.16 – **Sensibilité (A), FDP (B), et ratio sensibilité sur FDP (C) pour les 26 cancers de TCGA étudiés obtenus avec la pénalité elastic net - données simulées.**
Pour chaque cancer, la sensibilité et la proportion de fausses découvertes (FDP) sont calculées par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100$), et la médiane est calculée. La droite d'équation $y = x$ est tracée en pointillée.
FDP : taux de fausses découvertes (*False Discovery Proportion*).

2.6.4 Comparaison des méthodes de pénalisation

Les performances de sélection des pénalisations elastic net et adaptive elastic net sont comparables à la fois en terme de sensibilité et de taux de fausses découvertes (Fig. 2.17 B et E). En revanche, la pénalisation lasso obtient une sensibilité plus faible que les deux autres méthodes (Fig. 2.17 A et C), mais un taux de fausse découvertes plus faible (Fig. 2.17 D et F). Comme nous l'avons vu dans la partie 2.2.4, la pénalisation lasso sélectionne moins de gènes, et cela peut expliquer les différences que nous observons ici.

Ensuite, les performances de sélection obtenues sont mauvaises pour les trois pénalisation lasso, elastic net et adaptive elastic net (Fig. 2.17). Pour elastic net, les plus grandes

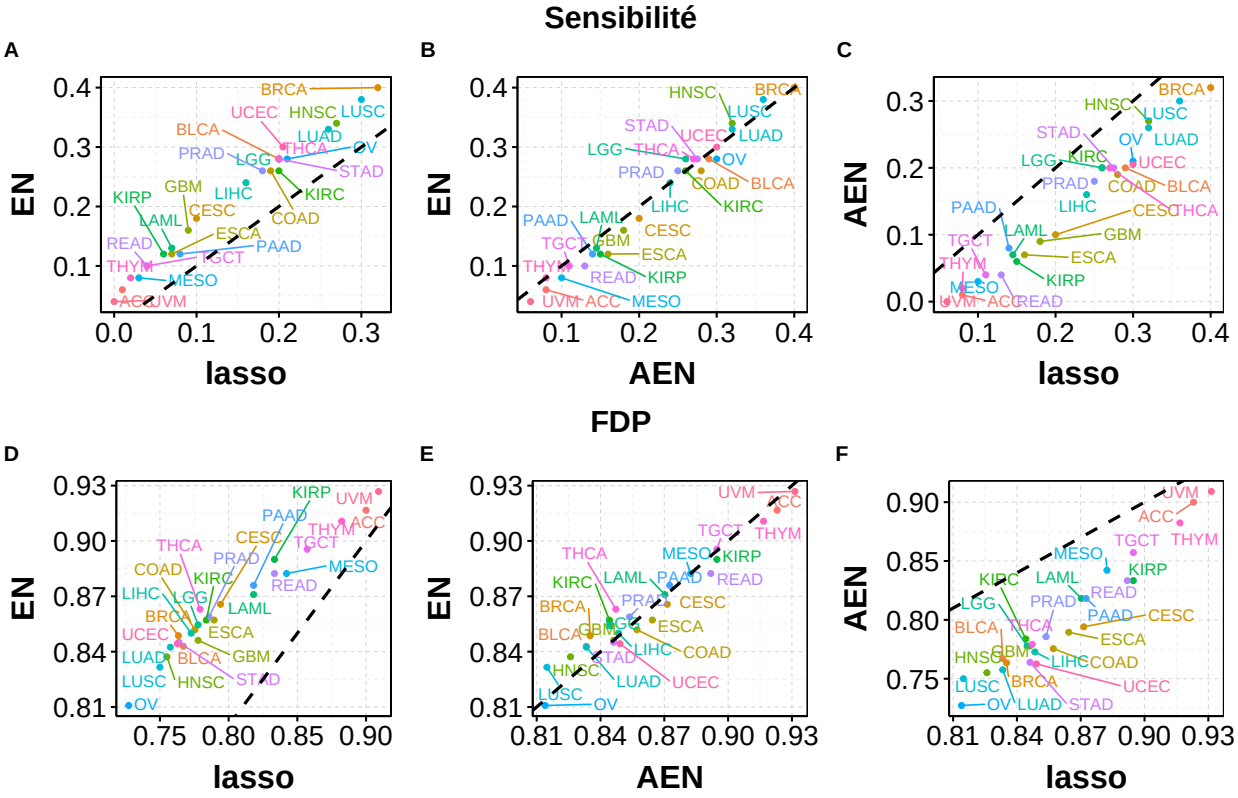


FIGURE 2.17 – **Comparaison des sensibilités (A, B, C) et des FDP (D, E, F) médians obtenus avec les pénalisations lasso, elastic net et adaptive elastic net pour les 26 cancers de TCGA étudiés - données simulées.**

Pour chaque cancer, la sensibilité et la proportion de fausses découvertes (FDP) sont obtenues par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100$), et la médiane est calculée. La droite d'équation $y = x$ est tracée en pointillée.
EN : elastic net ; AEN : adaptive elastic net.

sensibilités médianes sont atteintes pour BRCA et LUSC (0,4), mais pour ces deux jeux de données, les taux de fausses découvertes sont de 0,85 et 0,83 respectivement (Fig. 2.18). Le taux médian de fausses découvertes le plus bas est observé pour OV, mais il demeure très élevé (0,81).

Enfin, notons que les performances de sélection obtenues sur les données simulées diffèrent suivant les cancers (Fig. 2.17 et Fig. 2.18). Cela peut s'expliquer par les différents paramètres utilisés pour simuler les temps de survie (*i.e.* taux de censure, paramètres de la loi de Weibull, partie 2.4.1), par des structures de corrélations diverses des jeux de données d'expression génétique, ou par les différentes tailles d'échantillons suivant les cancers. De plus, remarquons qu'il existe une corrélation négative entre la sensibilité et le FDP (corrélation de -0,83, p-valeur < 0.001 - test de corrélation de Spearman, Fig. 2.18) : plus la sensibilité est élevé, plus le FDP est faible.

Dans la partie suivante, nous allons étudier l'influence du nombre de patients sur les performances de sélection.

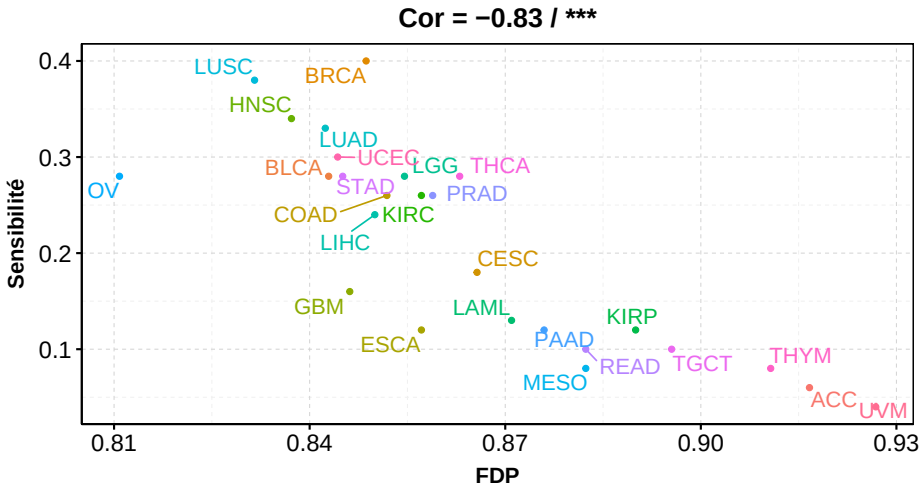


FIGURE 2.18 – **Sensibilité en fonction du taux de fausses découvertes sur les 26 cancers de TCGA étudiés et la pénalisation elastic net - données simulées.**
Pour chaque cancer, la sensibilité et la proportion de fausses découvertes (FDP) sont obtenus par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1;0,25;0,5;1$ et $N = 10;25;50;100$), et la médiane est calculée. La corrélation et le niveau de significativité d'un test de corrélation de Pearson est indiqué en haut du graphique (*** : p-valeur < 0,001).

2.6.5 Influence du nombre de patients

La sensibilité, la proportion de fausses découvertes et le nombre de gènes sélectionnés sont corrélés au nombre de patients dans le jeu de données. Ces trois corrélations sont significatives au niveau 0,01, et sont de 0,88, -0,58 et 0,92 respectivement. Plus le nombre de patients augmentent, plus la sensibilité et le nombre de gènes sélectionnés augmentent, et plus la proportion de fausses découvertes diminue. Ainsi, comme attendu, plus le nombre de patients est élevé, meilleures sont les performances de sélection.

La taille des cohortes de patients est limitée par des contraintes cliniques et pratiques, mais aussi par le coût du séquençage [MORTAZAVI et collab., 2008; SIMS et collab., 2014]. Une manière de remédier à cette dernière limitation est de diminuer la profondeur de séquençage pour réduire les coûts et augmenter la taille des cohortes (MILANEZ-ALMEIDA et collab. [2020], partie 1.3). L'étude de l'impact de la profondeur de séquençage des données miRNA-seq sur les capacités de prédiction fera l'objet du chapitre 4.

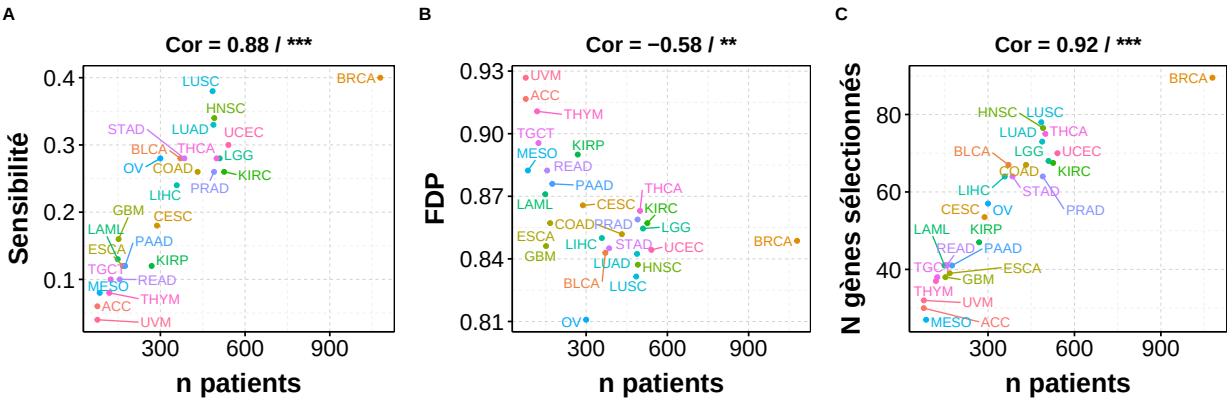


FIGURE 2.19 – **Sensibilité (A), FDP (B), et nombre de gènes sélectionnés (C) en fonction du nombre de patients dans le jeu d'apprentissage sur les 26 cancers de TCGA étudiés et la pénalisation elastic net - données simulées.**

Pour chaque cancer, la sensibilité et la proportion de fausses découvertes (FDP) sont obtenus par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1;0,25;0,5;1$ et $N = 10;25;50;100$), et la médiane est calculée. La corrélation et le niveau de significativité d'un test de corrélation de Pearson sont indiqués en haut de chaque graphique (** : p-valeur < 0,01 ; *** : p-valeur < 0,001).

2.7 Influence des performances de sélection sur la prédiction

Dans la partie précédente, nous avons vu que les performances de sélection sont décevantes pour les pénalisations lasso, elastic net et adaptive elastic net, à la fois en terme de stabilité des gènes sélectionnés, de sensibilité et de taux de fausses découvertes. En revanche, EIN-DOR et collab. [2005] ont montré que différents sous-ensembles de 70 covariables ne contenant aucun gène en commun prédisent de manière équivalente le statut des œstrogènes (positif ou négatif) dans le cancer du sein. Pour cela, les auteurs ont construit des scores de risque grâce à une régression logistique en utilisant successivement des groupes de 70 gènes classés par ordre croissant de leur p-valeur issue d'une régression logistique univariée. Des signatures qui ne partagent aucun gène peuvent donc aboutir à des performances de prédiction similaires dans cet exemple. Les mauvaises capacités de sélection n'impliquent donc pas nécessairement que les prédictions seront mauvaises.

Dans le cadre de notre étude, nous calculons les C-index obtenus d'une part avec l'ensemble des 1 000 gènes qui ont les plus faibles p-valeurs dans le modèle de Cox univarié (ensemble des prédicteurs), et d'autre part uniquement avec les N gènes qui composent la vérité terrain. Ainsi, dans le second cas, seuls les gènes effectivement corrélés à la survie (*i.e.* gènes avec des coefficients β_j non nuls dans les simulations) sont utilisés comme prédicteurs. La proportion de faux positifs est donc nulle, et un sous-ensemble de la vérité terrain est sélectionné par les méthodes lasso, elastic net et adaptive elastic net. Ce cas favorable peut être considéré comme un Oracle moins puissant que celui que nous

avons défini à la partie 2.4.3. Pour éviter toute confusion, nous parlerons ici de « prédictions obtenues si la vérité terrain était connue ». Comme dans la partie précédente, nous excluons le cas N = 1 000 des scénarios de simulations. Ainsi, les paramètres que nous utilisons pour les simulations sont $\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100$.

Pour tous les cancers et toutes les méthodes, le C-index médian serait significativement augmenté si la vérité terrain était connue par avance (p-valeur < 0.001, test des rangs signés de Wilcoxon unilatéral avec correction de Benjamini-Hochberg, Tab. 2.5a). Par exemple, pour elastic net et LGG, connaître la vérité terrain permet d'augmenter le C-index médian de 0,033 (Tab. 2.5a et Fig. 2.20.B). En moyenne suivant les 26 cancers étudiés et toujours pour elastic net, le taux important de fausses découvertes et la grande dimension engendre une diminution du C-index médian de 0,048.

En revanche, même dans le cas où la vérité terrain est connue par avance (*i.e.* proportion nulle de fausses découvertes), les performances oracles ne sont pas atteintes (Tab. 2.5b et Fig. 2.20.C). Les écarts entre les C-index oracles et estimés si la vérité terrain était connue sont significatifs pour tous les cancers et toutes les méthodes de pénalisation (p-valeur < 0.001, test des rangs signés de Wilcoxon unilatéral avec correction de Benjamini-Hochberg, Tab. 2.5b). Par exemple, pour elastic net et LGG, l'écart médian entre les C-index oracles et les C-index calculés avec uniquement les gènes de la vérité terrain est de 0.013 (Tab. 2.5b et Fig. 2.20.C). En moyenne pour l'ensemble des cancers, cette différence est de 0,031. Ces différences s'expliquent par une mauvaise estimation des indices pronostiques et donc des coefficients β_j dans le modèle de Cox.

Pour résumer, l'écart entre les prédictions obtenues avec le modèle de Cox pénalisé et les prédictions oracles que nous avons observé à la partie 2.5.3 s'explique à la fois par une proportion de faux positifs élevé, la grande dimension, et une mauvaise estimation du vecteur de coefficients $\boldsymbol{\beta}$ dans le modèle de Cox pénalisé.

Des résultats similaires sont obtenus en prenant la p-valeur du modèle de Cox et l'IBS comme métriques d'évaluation des prédictions (Fig. Annexe A.3 et A.4). Ainsi, diminuer le taux de fausses découvertes aiderait à augmenter significativement les performances de prédiction du modèle de Cox pénalisé.

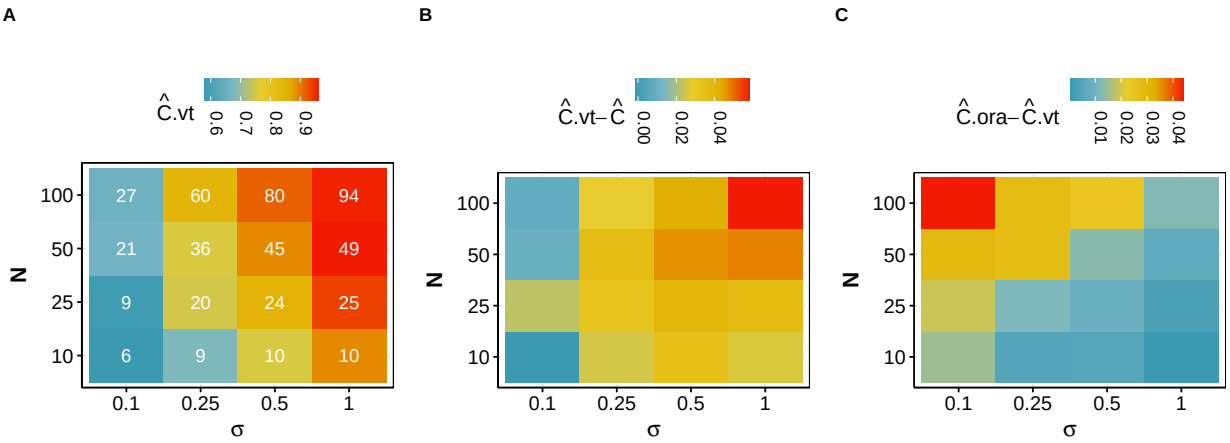


FIGURE 2.20 – **C-index médians obtenus si la vérité terrain était connue (A), et différences médianes avec les C-index obtenus avec l'ensemble des gènes (B), et avec les C-index oracles (C) pour elastic net et LGG - données simulées.**

Les C-index sont obtenus par 10 répétitions d'une validation croisée (K=5) pour chaque paramètre de simulation ($\sigma = 0,1;0,25;0,5;1$ et $N = 10;25;50;100$), et la médiane est calculée. Le jeu de données contient uniquement les 1 000 gènes qui ont les plus petites p-valeurs du modèle de Cox univarié.

$\hat{C}$.vt : C-index estimé par le modèle de Cox pénalisé avec uniquement les N gènes de la vérité terrain ; $\hat{C}$: C-index estimé par le modèle de Cox pénalisé avec les 1 000 gènes qui ont les plus petites p-valeurs du modèle de Cox univarié ; $\hat{C}$.ora : C-index calculé avec les indices pronostiques oracles utilisés pour les simulations. Le nombre médian de gènes sélectionnés par la pénalisation elastic net sur l'ensemble des N gènes de la vérité terrain est indiqué en blanc pour chaque paramètre de simulation dans le graphique A.

Cancer	ridge	lasso	EN	AEN
ACC	0,024	0,026	0,038	0,036
BLCA	0,074	0,041	0,05	0,043
BRCA	0,076	0,026	0,034	0,032
CESC	0,074	0,056	0,064	0,057
COAD	0,08	0,047	0,054	0,047
ESCA	0,091	0,065	0,078	0,067
GBM	0,065	0,053	0,064	0,053
HNSC	0,075	0,032	0,043	0,036
KIRC	0,052	0,026	0,033	0,03
KIRP	0,055	0,048	0,052	0,046
LAML	0,044	0,047	0,054	0,045
LGG	0,05	0,028	0,033	0,03
LIHC	0,064	0,041	0,046	0,042
LUAD	0,072	0,035	0,043	0,039
LUSC	0,085	0,038	0,047	0,04
MESO	0,046	0,042	0,05	0,046
OV	0,1	0,051	0,065	0,052
PAAD	0,056	0,039	0,05	0,045
PRAD	0,06	0,035	0,042	0,039
READ	0,075	0,057	0,068	0,064
STAD	0,07	0,039	0,046	0,042
TGCT	0,039	0,037	0,046	0,036
THCA	0,062	0,031	0,041	0,035
THYM	0,031	0,027	0,033	0,032
UCEC	0,079	0,039	0,048	0,045
UVM	0	0,029	0,034	0,03

(a) **Différences médianes entre les C-index estimés avec les gènes de la vérité terrain et les C-index estimés avec l'ensemble des 1 000 gènes.**

Cancer	ridge	lasso	EN	AEN
ACC	0,062	0,081	0,063	0,066
BLCA	0,019	0,021	0,018	0,021
BRCA	0,01	0,0099	0,0085	0,01
CESC	0,03	0,039	0,032	0,038
COAD	0,02	0,024	0,02	0,023
ESCA	0,045	0,059	0,048	0,05
GBM	0,034	0,038	0,035	0,038
HNSC	0,015	0,015	0,013	0,015
KIRC	0,017	0,017	0,016	0,016
KIRP	0,037	0,05	0,042	0,045
LAML	0,038	0,047	0,039	0,045
LGG	0,014	0,014	0,013	0,014
LIHC	0,02	0,023	0,021	0,022
LUAD	0,017	0,017	0,015	0,018
LUSC	0,014	0,015	0,014	0,016
MESO	0,058	0,065	0,056	0,063
OV	0,019	0,026	0,022	0,025
PAAD	0,033	0,042	0,036	0,038
PRAD	0,018	0,018	0,016	0,018
READ	0,052	0,067	0,057	0,059
STAD	0,016	0,019	0,016	0,018
TGCT	0,055	0,068	0,058	0,064
THCA	0,017	0,016	0,014	0,017
THYM	0,059	0,064	0,059	0,065
UCEC	0,015	0,016	0,014	0,016
UVM	0,059	0,067	0,059	0,064

(b) **Différences médianes entre les C-index oracles et les C-index estimés avec les gènes de la vérité terrain.**

TABLEAU 2.5 – **Comparaison des C-index oracles, estimés, et estimés avec uniquement les gènes de la vérité terrain pour les 26 cancers de TCGA - données simulées.**
Les 1 000 gènes utilisés sont ceux qui ont les plus faibles p-valeurs du modèle de Cox univarié dans les données réelles de TCGA. Nous calculons les médianes sur l'ensemble des simulations ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$).
EN : elastic net ; AEN : adaptive elastic net.

2.8 Conclusions

Tout d'abord, nous avons montré que les prédictions obtenues avec λ_{min} sont meilleures que celles obtenues avec λ_{1se}, et que les résultats sont très variables suivant les cancers. Ensuite, les différentes méthodes obtiennent globalement des résultats de performances comparables, avec de petites différences selon les cancers. En moyenne, la pénalisation ridge permet d'obtenir de meilleures prédictions. De plus, parmi les méthodes qui permettent de sélectionner un sous-ensemble de gènes, la pénalisation elastic net obtient en moyenne les meilleures performances. Cependant, les résultats obtenus dépendent fortement du jeu de données étudiés et ne peuvent être généralisés à l'ensemble des cancers. De plus, nous avons démontré l'apport des données mRNA-seq par rapport aux données cliniques classiques pour la prédiction pour 14 des 26 cancers étudiés.

Nous nous sommes ensuite intéressés aux performances de sélection des pénalisations lasso, elastic net, et adaptive elastic net. Nous avons tout d'abord montré que l'ensemble des gènes sélectionnés est très instable et dépend fortement des patients qui composent le jeu de données. Ensuite, nous avons mis en place une procédure de simulation avec des paramètres aussi proches des jeux de données étudiés que possible. Cette démarche nous permet d'évaluer les pénalisations du modèle de Cox dans des conditions favorables (*i.e.* simulations avec un modèle à risque proportionnel), à la fois en terme de sélection et de prédiction.

Ainsi, nous avons tout d'abord montré que les performances « oracles » de prédiction que l'on peut attendre du modèle de Cox (*i.e.* prédictions obtenues si le vecteur $\boldsymbol{\beta}$ de coefficients utilisés pour les simulations était connu) ne sont pas atteintes. Par exemple, l'écart médian que nous avons observé entre les C-index oracles et estimés est de 0.088 pour elastic net et en moyenne pour l'ensemble des 26 cancers étudiés. Ensuite, notre processus de simulation démontre que les performances de sélection sont mauvaises pour les trois pénalisations (*i.e.* lasso, elastic net et adaptive elastic net). Pour l'ensemble des cancers et des méthodes, la sensibilité ne dépasse pas 0,40, et la proportion de fausses découvertes est toujours supérieur à 0,70. Ces performances limitées sont dues à la complexité de la tâche : grand nombre de prédicteurs (la quantité normalisée des transcrits) par rapport au nombre de patients, et la structure de corrélation de ces transcrits qui dépend de chaque cancer. Cela donne une borne supérieure des performances que nous pouvons attendre sur données réelles avec le modèle de Cox qui suppose les risques proportionnels.

Finalement, afin d'étudier l'impact de ces mauvaises performances de sélection sur la prédiction, nous avons comparé les prédictions obtenues si la vérité terrain était connue par avance, et celles obtenues avec l'ensemble des gènes. Nous avons ainsi pu montrer que le taux élevé de faux positifs a un impact négatif important sur la qualité des prédictions. En revanche, même dans le cas où la vérité terrain est connue, les prédictions oracles ne sont pas atteintes. Ainsi, l'écart entre les prédictions oracles et estimées est dû

à la fois au taux important de faux positifs, à la grande dimension, et à des biais dans l'estimation des coefficients β_j et des indices pronostiques.

Le pré-filtrage des gènes permet de retenir de manière univarié qu'un sous-ensemble de gènes qui serviront ensuite de prédicteurs dans le modèle de Cox multivarié. Cette méthodologie permet de réduire simplement la dimension et à montrer des résultats prometteurs en pratique [BØVELSTAD et collab., 2007; MILLER et collab., 2002]. Ainsi, dans le chapitre suivant, nous nous intéresserons à l'impact de méthodes de pré-filtrage, permettant de diminuer la dimension, sur les performances de prédiction du modèle de Cox pénalisé.

2.9 Références

ARAN, D. et collab.. 2015, «Systematic pan-cancer analysis of tumour purity», *Nature Communications*, vol. 6, nº 1, doi :10.1038/ncomms9971, p. 8971, ISSN 2041-1723. URL https://www.nature.com/articles/ncomms9971, number : 1 Publisher : Nature Publishing Group. 61

ASSEL, M. et collab.. 2017, «The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models», *Diagnostic and Prognostic Research*, vol. 1, nº 1, doi : 10.1186/s41512-017-0020-3, p. 19, ISSN 2397-7523. URL https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-017-0020-3. 50

BENDER, R. et collab.. 2005, «Generating survival times to simulate Cox proportional hazards models», *Statistics in Medicine*, vol. 24, nº 11, doi :10.1002/sim.2059, p. 1713–1723, ISSN 02776715, 10970258. URL http://doi.wiley.com/10.1002/sim.2059. 61

BENNER, A. et collab.. 2010, «High-dimensional Cox models : the choice of penalty as part of the model building process», *Biometrical Journal. Biometrische Zeitschrift*, vol. 52, nº 1, doi :10.1002/bimj.200900064, p. 50–69, ISSN 1521-4036. 47

BRIERLEY, J. D. et collab.. 2017, *TNM Classification of Malignant Tumours, 8th Edition*, Wiley. 57

BØVELSTAD, H. M. et collab.. 2007, «Predicting survival from microarray data–a comparative study», *Bioinformatics (Oxford, England)*, vol. 23, nº 16, doi :10.1093/bioinformatics/btm305, p. 2080–2087, ISSN 1367-4811. 47, 85

BØVELSTAD, H. M. et collab.. 2009, «Survival prediction from clinico-genomic models - a comparative study», *BMC Bioinformatics*, vol. 10, nº 1, doi :10.1186/1471-2105-10-413, p. 413, ISSN 1471-2105. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-413. 56

DOMANY, E. 2014, «Using High-Throughput Transcriptomic Data for Prognosis : A Critical Overview and Perspectives», *Cancer Research*, vol. 74, nº 17, doi :10.1158/0008-5472.CAN-13-3338, p. 4612–4621, ISSN 0008-5472, 1538-7445. URL http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-13-3338. 74

EIN-DOR, L. et collab.. 2005, «Outcome signature genes in breast cancer : is there a unique set?», *Bioinformatics*, vol. 21, nº 2, doi :10.1093/bioinformatics/bth469, p. 171–178, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/21/2/171/186749, publisher : Oxford Academic. 74, 80

HARRELL, F. E. et collab.. 1996, «Multivariable Prognostic Models : Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors», *Statistics in Medicine*, vol. 15, nº 4, doi :10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4, p. 361–387, ISSN 1097-0258. URL https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4. 50

HERRMANN, M. et collab.. 2020, «Large-scale benchmark study of survival prediction methods using multi-omics data», *Briefings in Bioinformatics*, doi :10.1093/bib/bbaa167, p. bbaa167, ISSN 1467-5463, 1477-4054. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa167/5895463. 56, 59

JIANG, Y. et collab.. 2016, «Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis», *Genomics*, vol. 107, nº 6, doi :10.1016/j.ygeno.2016.04.005, p. 223–230, ISSN 1089-8646. 61

KALBFLEISCH, J. D. et R. L. PRENTICE. 2011, *The Statistical Analysis of Failure Time Data*, AMLBook, ISBN 9781118032985, doi :10.1002/9781118032985. 61

LU, J. et collab.. 2018, «Clinical significance of prognostic score based on age, tumor size, and grade in gastric cancer after gastrectomy», *Cancer Management and Research*, vol. 10, doi :10.2147/CMAR.S171663, p. 4279–4286, ISSN 1179-1322. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6183590/. 57

LÓPEZ DE MATURANA, E., L. ALONSO, P. ALARCÓN, I. A. MARTÍN-ANTONIANO, S. PINEDA, L. PIORNO, M. L. CALLE et N. MALATS. 2019, «Challenges in the Integration of Omics and Non-Omics Data», *Genes*, vol. 10, nº 3, doi :10.3390/genes10030238, ISSN 2073-4425. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6471713/. 56

LÓPEZ DE MATURANA, E., S. PINEDA, A. BRAND, K. VAN STEEN et N. MALATS. 2016, «Toward the integration of Omics data in epidemiological studies : still a "long and winding road"», *Genetic Epidemiology*, vol. 40, nº 7, doi :10.1002/gepi.21992, p. 558–569, ISSN 1098-2272. 59

MEINSHAUSEN, N. et P. BÜHLMANN. 2010, «Stability selection», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 72, nº 4, doi : 10.1111/j.1467-9868.2010.00740.x, p. 417–473, ISSN 1467-9868. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x, _eprint : https ://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x. 75

METZELER, K. H. et collab.. 2008, «An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia», *Blood*, vol. 112, nº 10, doi : 10.1182/blood-2008-02-134411, p. 4193–4201, ISSN 1528-0020. 47

MICHIELS, S. et collab.. 2005, «Prediction of cancer outcome with microarrays : a multiple random validation strategy», *The Lancet*, vol. 365, nº 9458, doi : 10.1016/S0140-6736(05)17866-0, p. 488–492, ISSN 0140-6736. URL http://www.sciencedirect.com/science/article/pii/S0140673605178660. 74

MILANEZ-ALMEIDA, P. et collab.. 2020, «Cancer prognosis with shallow tumor RNA sequencing», *Nature Medicine*, vol. 26, nº 2, doi :10.1038/s41591-019-0729-3, p. 188–192, ISSN 1078-8956, 1546-170X. URL http://www.nature.com/articles/s41591-019-0729-3. 56, 57, 59, 60, 79

MILLER, L. D. et collab.. 2002, «Optimal gene expression analysis by microarrays», *Cancer Cell*, vol. 2, nº 5, doi :10.1016/S1535-6108(02)00181-2, p. 353–361, ISSN 1535-6108. URL http://www.sciencedirect.com/science/article/pii/S1535610802001812. 85

MOOD, A. M. 1974, *Introduction to the theory of statistics*, McGraw-Hill, New York, NY, US. Pages : xiii, 433. 61

MORTAZAVI, A. et collab.. 2008, «Mapping and quantifying mammalian transcriptomes by RNA-Seq», *Nature Methods*, vol. 5, nº 7, doi :10.1038/nmeth.1226, p. 621–628, ISSN 1548-7091, 1548-7105. URL http://www.nature.com/articles/nmeth.1226. 79

OJEDA, F. M. et collab.. 2016, «Comparison of Cox Model Methods in A Low-dimensional Setting with Few Events», *Genomics, Proteomics & Bioinformatics*, vol. 14, nº 4, doi :10.1016/j.gpb.2016.03.006, p. 235–243, ISSN 1672-0229. URL http://www.sciencedirect.com/science/article/pii/S1672022916300390. 47

RICKETTS, C. J. et collab.. 2018, «The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma», *Cell Reports*, vol. 23, nº 1, doi :10.1016/j.celrep.2018.03.075, p. 313–326.e5, ISSN 2211-1247. URL https://www.cell.com/cell-reports/abstract/S2211-1247(18)30436-4, publisher : Elsevier. 59

ROBERTS, S. et G. NOWAK. 2014, «Stabilizing the lasso against cross-validation variability», *Computational Statistics & Data Analysis*, vol. 70, doi :10.1016/j.csda.2013.09.

008, p. 198–211, ISSN 0167-9473. URL http://www.sciencedirect.com/science/article/pii/S016794731300323X. 47

ROSENBERG, J. et collab.. 2005, «The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database», *Breast Cancer Research and Treatment*, vol. 89, nº 1, doi :10.1007/s10549-004-1470-1, p. 47–54, ISSN 0167-6806. 57

SIMS, D. et collab.. 2014, «Sequencing depth and coverage : key considerations in genomic analyses», *Nature Reviews Genetics*, vol. 15, nº 2, doi :10.1038/nrg3642, p. 121–132, ISSN 1471-0056, 1471-0064. URL http://www.nature.com/articles/nrg3642. 79

TERNÈS, N. et collab.. 2017, «Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces», *Biometrical Journal. Biometrische Zeitschrift*, vol. 59, nº 4, doi :10.1002/bimj.201500234, p. 685–701, ISSN 1521-4036. 59

THERNEAU, T. M. 2020, *A Package for Survival Analysis in R*. URL https://CRAN.R-project.org/package=survival, r package version 3.2-3. 57

THIESE, M. S. et collab.. 2016, «P value interpretations and considerations», *Journal of Thoracic Disease*, vol. 8, nº 9, doi :10.21037/jtd.2016.08.16, p. E928–E931, ISSN 2072-1439. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059270/. 52

VAN 'T VEER, L. J. et collab.. 2002, «Gene expression profiling predicts clinical outcome of breast cancer», *Nature*, vol. 415, nº 6871, doi :10.1038/415530a, p. 530–536, ISSN 0028-0836. 47

VOLKMANN, A., R. DE BIN, W. SAUERBREI et A.-L. BOULESTEIX. 2019, «A plea for taking all available clinical information into account when assessing the predictive value of omics data», *BMC Medical Research Methodology*, vol. 19, nº 1, doi :10.1186/s12874-019-0802-0, p. 162, ISSN 1471-2288. URL https://doi.org/10.1186/s12874-019-0802-0. 56, 58, 59

WAN, F. 2017, «Simulating survival data with predefined censoring rates for proportional hazards models», *Statistics in Medicine*, vol. 36, nº 5, doi :10.1002/sim.7178, p. 838–854, ISSN 1097-0258. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7178, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7178. 62

WEI, H. et collab.. 2017, «MiR-638 inhibits cervical cancer metastasis through Wnt/-catenin signaling pathway and correlates with prognosis of cervical cancer patients», *European Review for Medical and Pharmacological Sciences*, vol. 21, nº 24, doi :10.26355/eurrev_201712_13999, p. 5587–5593, ISSN 2284-0729. 59

ZHAO, Q. et collab.. 2015, «Combining multidimensional genomic measurements for predicting cancer prognosis : observations from TCGA», *Briefings in Bioinformatics*, vol. 16, nº 2, doi :10.1093/bib/bbu003, p. 291–303, ISSN 1467-5463. URL https://academic.oup.com/bib/article/16/2/291/246070, publisher : Oxford Academic. 56, 61

ZHU, B., N. SONG, R. SHEN, A. ARORA, M. J. MACHIELA, L. SONG, M. T. LANDI, D. GHOSH, N. CHATTERJEE, V. BALADANDAYUTHAPANI et H. ZHAO. 2017, «Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers», *Scientific Reports*, vol. 7, nº 1, doi :10.1038/s41598-017-17031-8, p. 16 954, ISSN 2045-2322. URL https://www.nature.com/articles/s41598-017-17031-8, number : 1 Publisher : Nature Publishing Group. 56

ZOU, H. et T. HASTIE. 2005, «Regularization and variable selection via the elastic net», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 67, nº 2, doi :10.1111/j.1467-9868.2005.00503.x, p. 301–320, ISSN 1369-7412, 1467-9868. URL http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x. 53, 75

Chapitre 3

Pré-filtrage univarié des données d'expression génétiques et prédiction avec le modèle de Cox multivarié

Sommaire

3.1 Pré-filtrage univarié des données d'expression génétique 92
3.1.1 Intérêt du pré-filtrage 92
3.1.2 Pré-filtrage supervisé 93
3.1.3 Pré-filtrage non-supervisé 93
3.1.4 Objectif du chapitre 95
3.1.5 Choix des cancers étudiés 96
3.2 Impact du pré-filtrage sur les capacités de prédiction du modèle de Cox pénalisé 96
3.2.1 Choix des seuils pour les deux méthodes de pré-filtrage 96
3.2.2 Impact du pré-filtrage sur les performances de prédiction 99
3.3 Comparaison des méthodes de pénalisation du modèle de Cox après pré-filtrage 102
3.4 Comparaison du pré-filtrage bi-dimensionnel à l'algorithme *Iterative Sure Independance Screening* 102
3.4.1 L'algorithme *Iterative Sure Independance Screening* 102
3.4.2 Comparaison avec le pré-filtrage bi-dimensionnel 104
3.5 Conclusions et perspectives 106
3.6 Références 107

3.1 Pré-filtrage univarié des données d'expression génétique

3.1.1 Intérêt du pré-filtrage

Comme nous l'avons souligné en introduction (partie 1.8.2), une première étape de pré-filtrage univarié des gènes peut permettre d'obtenir de meilleures prédictions avec le modèle de Cox pénalisé BENNER et collab. [2010]. Cette stratégie permet de faire face à la « malédiction de la grande dimension » en réduisant de manière simple et intuitive le nombre p de prédicteurs (partie 1.8.1). Elle consiste à assigner un score à chaque gène, et à garder uniquement les gènes dont les scores sont supérieurs (ou inférieurs) à un certain seuil. Dans ce chapitre, nous allons explorer deux méthodes de pré-filtrage univarié basées sur des scores distincts :

— la p-valeur d'un test d'association entre le niveau d'expression du gène et la survie.

— la variabilité du gène entre les patients.

Pour commencer, filtrer les gènes dont l'association avec la survie est faible permet de réduire facilement la dimension en n'incluant dans le modèle multivarié que les gènes qui sont corrélés à la survie et qui ont donc a priori un intérêt pour la prédiction BØVELSTAD et collab. [2007].

Ensuite, les gènes peu variables parmi les patients sont plus sujets aux problèmes techniques liés au séquençage et à un mauvais rapport signal-à-bruit. En effet, de faibles variations montrent que le signal potentiellement utile pour séparer les patients, ou prédire la survie, est probablement noyé dans le bruit de mesure. De plus, les gènes peu variables sont moins susceptibles de séparer correctement les patients en terme de risque [MILLER et collab., 2002].

Ainsi, le pré-filtrage bidimensionnel permet de garantir que les gènes utilisés dans le modèle de Cox sont pertinents pour prédire la survie, et suffisamment variables parmi les patients pour minimiser l'influence du bruit de mesure. Cette réduction de la dimension de manière univariée permet d'atténuer les problèmes de sur-apprentissage liés à la « malédiction de la grande dimension » (partie 1.8.1). Nous avons vu que les gènes sélectionnés par le modèle de Cox pénalisé sont très instables pour lasso, elastic net et adaptive elastic net. Ainsi, réduire la dimension dans une première étape de pré-filtrage univarié peut permettre d'obtenir une plus grande stabilité des gènes qui composent le score de risque, et éviter des confusions dans l'interprétation lié à cette instabilité [BOULESTEIX et collab., 2020; DOMANY, 2014].

Cependant, ces deux méthodes de pré-filtrage ne prennent pas en compte la structure de corrélation entre les données, et BØVELSTAD et collab. [2007] ont montré que les prédictions obtenues avec un ensemble de gènes sélectionnés par le modèle de Cox univarié sont moins bonnes que celles obtenues directement avec le modèle de Cox pénalisé (multivarié). Ainsi, le choix des seuils de pré-filtrage apparaît essentiel pour optimiser les prédictions.

3.1.2 Pré-filtrage supervisé

La première méthode repose sur le modèle de Cox univarié. Elle consiste à assigner une p-valeur à chaque gène afin de tester l'association entre le niveau d'expression du gène et la survie dans le jeu de données d'apprentissage. Ce pré-filtrage est dit « supervisé » car les données de survie sont utilisées en plus des données génétiques. L'hypothèse nulle du test pour le gène j est « $H_0^j : \beta_j = 0$ », où β_j est le coefficient du modèle de Cox univarié : $h(t;X_j) = h_0(t)\exp(\beta_j X_j)$. La p-valeur est calculée grâce à un test du rapport de vraisemblance (partie 1.7.4). RAMAN et collab. [2019] ont montré que le modèle de Cox univarié est le plus performant parmi un ensemble de huit méthodes classiques pour sélectionner des gènes associés à la survie. Cette méthode est la fois utilisée pour étudier l'association d'un gène avec la survie [DAI et collab., 2017; SHAO et collab., 2019], ou pour pré-filtrer de manière univarié l'ensemble des prédicteurs avant d'utiliser un modèle multivarié [JIANG et collab., 2016; ZHAO et collab., 2015]. Dans notre étude, les p-valeurs sont corrigées par le méthode de Benjamini-Hochberg (partie 1.5.3, BENJAMINI et HOCHBERG [1995]), et seuls les gènes qui ont une p-valeur corrigée inférieure à un certain seuil sont retenus dans le modèle de Cox pénalisé.

Pour éviter toute confusion, notons ici que le modèle de Cox univarié a déjà été présenté au paragraphe 1.7.4, mais dans l'optique d'évaluer les capacités de prédiction d'un modèle en étudiant l'association entre les indices pronostiques du jeu données test et les temps de survie. L'utilisation du modèle de Cox univarié est tout autre dans le contexte du pré-filtrage : il est calculé pour chaque gène à partir des données du jeu d'apprentissage, et permet d'éliminer des gènes dont l'association univariée avec la survie n'est pas significative.

3.1.3 Pré-filtrage non-supervisé

La deuxième méthode consiste à filtrer les gènes dont le niveau d'expression n'est pas assez variable, ou dispersé, parmi les patients. Elle repose sur l'Écart Interquartile (EI) du niveau d'expression des gènes à travers les patients du jeu de données d'apprentissage. Seuls les gènes qui ont un EI supérieur à un certain seuil sont retenus dans le modèle de Cox pénalisé. Les données de survie ne sont pas utilisées pour le filtrage et la méthode est ainsi qualifiée de « non-supervisée ».

Cette méthode est classiquement utilisée sur les données de puces à ADN (partie 1.3.1) [FA et collab., 2019; LIAO et collab., 2011; MICHIELS et collab., 2005]. En effet, les gènes peu variables suivant les patients sont plus sujets aux bruits de mesure et sont moins susceptibles de distinguer les patients en terme de risque [MILLER et collab., 2002]). En outre l'EI a l'avantage d'être une mesure de dispersion plus robuste aux éventuelles données incorrectes ou aberrantes que d'autres statistiques descriptives classiques telles que la variance, ou l'écart-type, empirique.

Les données d'expression génétique TCGA que nous utilisons ont été générées par la

technique RNA-seq qui est basée sur une stratégie différente des puces à ADN (partie 1.3). Ces données d'expression sont issues d'un processus de comptage qui induit une relation entre la médiane et l'EI comme illustré sur la Figure 3.1.A. L'EI a alors tendance à augmenter avec le niveau médian d'expression des gènes. Cette caractéristique est typique du modèle de Poisson, dans lequel l'espérance est égale à la variance. Ainsi, ceci favorise les gènes en moyenne les plus exprimés dans notre méthode de filtrage, même si la variance inter-patients n'est pas significative. Nous observons aussi ce biais sur les données CPM (Fig. 3.1.B).

Ensuite, l'application du logarithme en base 2 permet de remédier au problème de l'asymétrie de la distribution des données [ZWIENER et collab., 2014], mais assigne un EI plus important aux gènes peu exprimés qu'aux gènes fortement exprimés (Fig. 3.1.C). Le biais décrit au paragraphe précédent est donc toujours observé, mais dans un sens opposé.

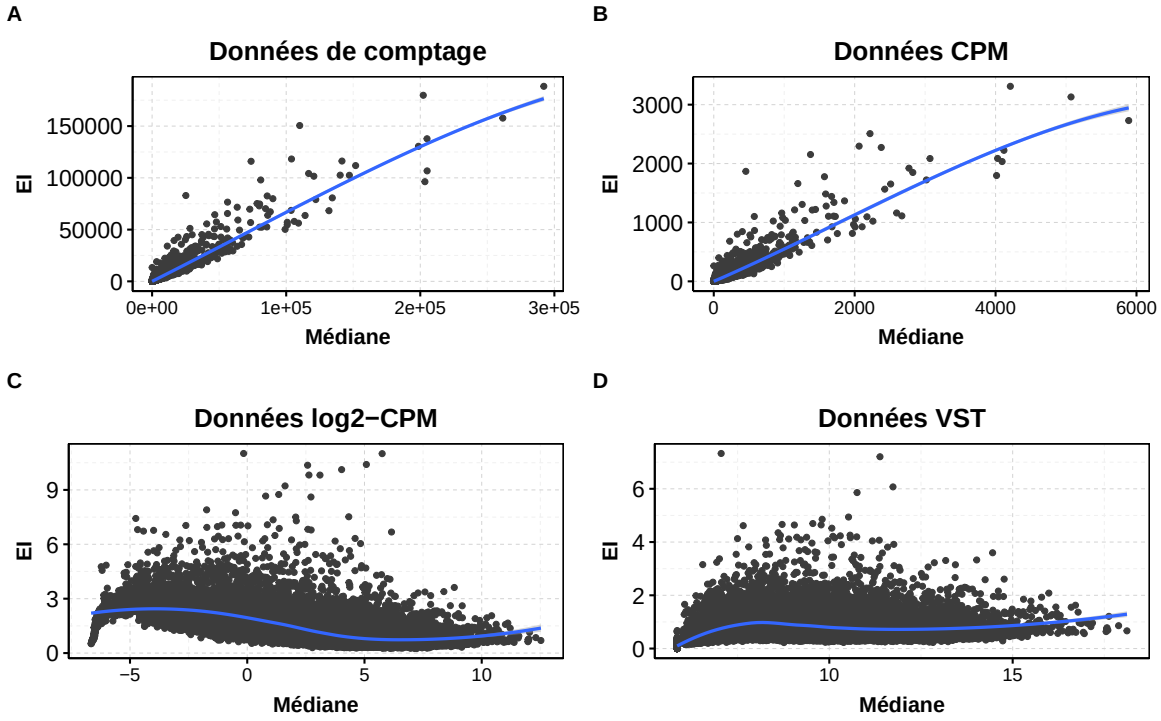


FIGURE 3.1 – **Relation entre l'écart interquartile et la médiane dans les données de comptage RNA-seq (A), CPM (B), log2-CPM (C), et VST (D) pour LGG.**
La courbe bleue est l'estimation *loess* calculée sur l'ensemble des points.

Finalement, pour remédier à ces biais, nous utilisons un algorithme de stabilisation de la variance [ANDERS et HUBER, 2010]. Le principe de cette méthode est d'appliquer une fonction g à chaque données CPM. Le prinicipe est le suivant. En notant X une variable aléatoire dont l'espérance et la variance sont liés par une fonction h tel que $E(X) = \mu$ et $Var(X) = k(\mu)$, un développement de Taylor au premier ordre permet d'obtenir la relation :

$$\begin{aligned} Y &= g(X) \\ &= g(\mu) + g'(\mu)(X - \mu) \end{aligned}$$

Cette nouvelle variable aléatoire Y a une variance de $Var(Y) = Var(X)g'(\mu)^2 = k(\mu)g'(\mu)^2$. En imposant que cette variance soit constante, on obtient :

$$g'(\mu) = \frac{a}{\sqrt{k(\mu)}},$$

avec a un réel constant. La transformation que nous appliquons aux données CPM est alors :

$$g(x) = \int_0^x \frac{1}{\sqrt{k(\mu)}} d\mu,$$

La fonction k est définie comme une fonction polynomiale du second degré qui relie la variance à l'espérance, et les paramètres sont estimés à partir des espérances et des variances de chaque gène. Les résultats obtenus pour LGG sont montrés dans le Figure 3.1.D. Cette méthode est implémentée dans la fonction `vst` (« *Variance Stabilizing Transformation* ») du package `DESeq2` [LOVE et collab., 2014].

Ainsi, pour chaque gène, nous faisons un pré-filtrage supervisé sur l'EI calculé sur les données *Variance Stabilizing Transformation* (VST) de l'ensemble des patients du jeu d'apprentissage.

3.1.4 Objectif du chapitre

Finalement, les scores suivants sont calculés sur l'ensemble des patients du jeu de données d'apprentissage et sont assignés à chaque gène pour effectuer un pré-filtrage univarié :

— les p-valeurs du modèle de Cox univarié avec correction de Benjamini-Hochberg.

— les écarts interquartiles (EI) de chaque gène après stabilisation de la variance.

Ces deux types de pré-filtrages sont souvent utilisés en pratique mais sans justification du seuil choisi, à la fois pour le pré-filtrage supervisé [JIANG et collab., 2016; ZHAO et collab., 2015] et non supervisé [FA et collab., 2019; LIAO et collab., 2011; MICHIELS et collab., 2005].

Ainsi, les objectifs de ce chapitre sont :

1. de proposer une méthode pour estimer les seuils optimaux à la fois pour le pré-filtrage supervisé et non-supervisé avant d'utiliser le modèle de Cox pénalisé.
2. d'étudier l'impact du pré-filtrage bidimensionnel (*i.e.* supervisé et non supervisé) sur la prédiction.

3. de comparer les pénalisations du modèle de Cox (*i.e.* ridge, lasso, elastic net, adaptive elastic net) après une première étape de pré-filtrage bidimensionnel.
4. de comparer différentes méthodes de pré-filtrage classiques.

Pour éviter le sur-apprentissage, il est important de noter que nous calculons ces scores avec les patients du jeu de données d'apprentissage [BOULESTEIX et collab., 2020].

3.1.5 Choix des cancers étudiés

Parmi les 26 cancers retenus en introduction (partie 1.5.2), nous avons choisi de ne retenir que les cancers pour lesquels les données transcriptomiques d'ARNm permettent de prédire correctement la survie dans le modèle de Cox avec pénalisation ridge. En effet, ce sont pour ces cancers que les données du niveau d'expression des ARNm ont le plus d'applications et de valorisations potentielles. Dans ce sens, nous considérons qu'un jeu de données a un pouvoir prédictif si le C-index médian est significativement supérieur à 0,6. Ainsi, nous avons calculé 50 C-index par 10 répétitions d'une validation croisée (K=5) pour chaque cancer avec les données mRNA-seq et un modèle de Cox pénalisé (ridge) (Fig. 1.6). Un test de Wilcoxon unilatéral nous permet alors de déterminer si sur la médiane m des C-index obtenus est significativement supérieure à 0,6. Ce test a pour hypothèse nulle $H_0 : m < 0,6$, et pour hypothèse alternative $H_1 : m > 0.6$. La méthode de Benjamini-Hochberg (partie 1.5.3, BENJAMINI et HOCHBERG [1995]) nous permet de corriger les p-valeurs obtenues pour les 26 cancers.

Suite à ces tests et au niveau 5%, nous avons sélectionné 16 cancers parmi les 26 (ACC, BLCA, BRCA, CESC, HNSC, KIRC, KIRP, LAML, LGG, LIHC, LUAD, MESO, PAAD, PRAD, UCEC, UVM). De plus, pour les 10 autres cancers, nous avons observé un nombre importants d'erreurs dans l'apprentissage du modèle de Cox pénalisé (partie 2.2.4). Enlever ces 10 cancers permet ainsi de nous prévaloir d'erreurs potentielles dans la détermination des seuils optimaux de pré-filtrage.

3.2 Impact du pré-filtrage sur les capacités de prédiction du modèle de Cox pénalisé

3.2.1 Choix des seuils pour les deux méthodes de pré-filtrage

Les différents seuils testés sont

$$p_{\text{seuil}} = 0,01\,;\,0,05\,;\,0,1\,;\,0,2\,;\,0,5\,;\,\mathbf{1}$$

pour la p-valeur du modèle de Cox univarié, et

$$\text{EI}_{\text{seuil}} = \mathbf{0}\,;\,0,5\,;\,1\,;\,1,5\,;\,2\,;\,2,5$$

pour l'EI. Les seuils indiqués en gras ne filtrent aucun gène (équivalent à aucun seuil). Ainsi, les gènes qui définissent l'ensemble des prédicteurs utilisés dans le modèle de Cox multivarié sont ceux qui ont à la fois une p-valeur du modèle de Cox inférieure à p_{seuil}, et un EI supérieur à EI_{seuil}.

Remarquons que les p-valeurs sont sensibles au nombre de patients. Ainsi, les seuils choisis pour la validation croisée ont une influence sur le nombre de gènes pré-filtrés de manière supervisée : pour deux jeux de données équivalents mais de tailles différentes, les p-valeurs du jeu de données qui a le nombre de patients le plus élevé auront tendance à être plus faible. Le choix des seuils peut donc être adapté en fonction de la dimension du jeu de données utilisé.

Ensuite, pour chaque seuil défini ci-dessus, nous retenons uniquement les gènes non-filtrés sur le jeu de données d'apprentissage, nous apprenons un modèle de Cox pénalisé sur le jeu d'apprentissage, et nous calculons les trois métriques d'évaluation des capacités de prédiction (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS) sur le jeu de données test. Nous répétons ce procédé 50 fois par 10 répétitions d'une validation croisée en K = 5 groupes (Fig. 1.6). Les seuils optimaux sont ceux qui maximisent (resp. minimisent) le C-index (resp. la p-valeur du modèle de Cox univarié, l'IBS) (Fig. 3.2).

Tout d'abord, les seuils estimés sont très différents suivant les cancers. Par exemple, pour la pénalisation elastic net et le C-index comme mesure de la qualité de prédiction, les seuils optimaux pour KIRP sont de 0,01 pour la p-valeur du modèle de Cox univarié et de 0 pour l'EI, alors qu'ils sont respectivement de 1 et de 2,5 pour PAAD (Fig. 3.3). Cette diversité se retrouve pour les deux autres métriques (*i.e.* p-valeur du modèle de Cox univarié et IBS, Fig. Annexe A.11) et les autres pénalisations (données non montrées).

Ensuite, le seuil optimal de l'EI calculé avec le C-index et la p-valeur du modèle de Cox univarié sont corrélés entre les cancers pour les pénalisations lasso, elastic net et adaptive elastic net (corrélation > 0,5 et p-valeur < 0,05, test de corrélation de Spearman, données non montrées). Ainsi, ces deux métriques se comportent de manière comparables pour choisir le seuil sur l'EI en optimisant les prédictions. En revanche, ces corrélations ne sont pas observées pour l'IBS (*i.e.* les seuils optimaux de l'EI choisis avec l'IBS ne sont pas corrélés à ceux choisis avec le C-index ou la p-valeur du modèle de Cox univarié). Cela s'explique par la moindre sensibilité de l'IBS observée au chapitre 2 (partie 2.2.3).

Ensuite, ces corrélations ne sont pas observées avec le seuil optimal de la p-valeur du modèle de Cox. En effet, le pré-filtrage semble plus efficace avec l'EI plutôt qu'avec la p-valeur du modèle de Cox univarié (stratification horizontale sur la Figure 3.2). Ainsi, le seuil optimal du filtrage sur la p-valeur du modèle de Cox univarié dépend fortement de la métrique d'évaluation choisie. De plus, pour tous les cancers sauf pour PAAD, un pré-filtrage sur l'EI est effectué dans le cas optimal, alors qu'aucun pré-filtrage n'est fait sur la p-valeur du modèle de Cox univarié pour 6 cancers (PAAD, LUAD, ACC, LIHC, MESO, PRAD).

De la même manière, le choix des seuils sur l'EI sont similaires pour les pénalisations

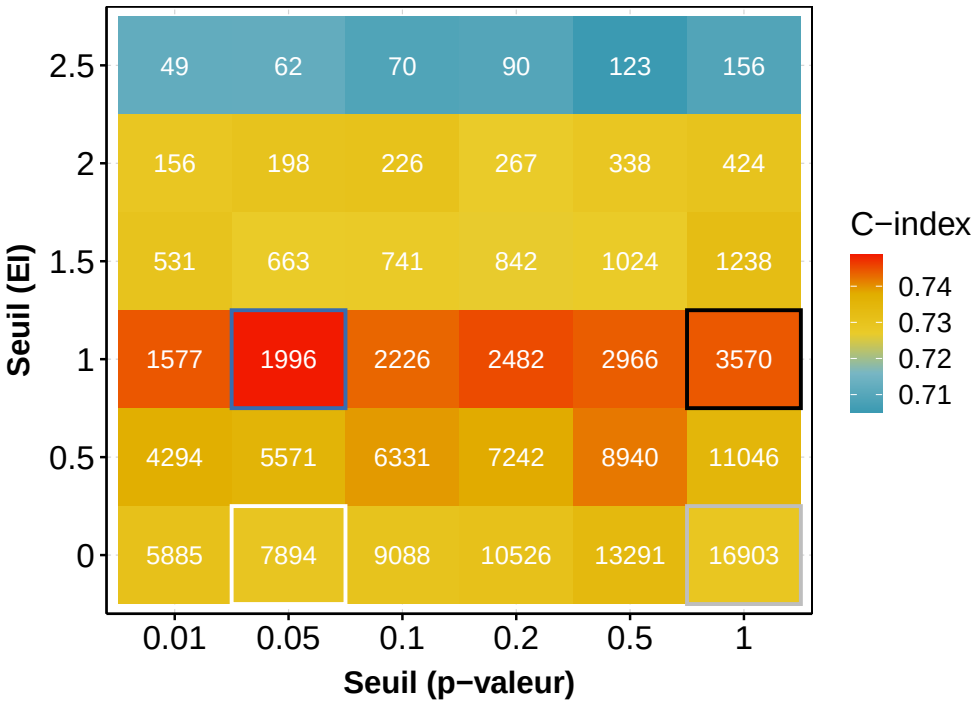


FIGURE 3.2 – **C-index médians obtenus pour différents seuils de pré-filtrage pour LGG.** Pour chaque seuil, les C-index sont calculés par 10 répétitions d'une validation croisée en K = 5 groupes, et le pré-filtrage des gènes est fait sur le jeu de données d'apprentissage. Dans chaque case, le chiffre indiqué en blanc correspond au nombre médian de gènes retenus après le pré-filtrage bi-dimensionnel.
Case entourée de noire : résultats obtenus avec uniquement un pré-filtrage sur l'EI ; case entourée de blanc : résultats obtenus avec uniquement un pré-filtrage sur la p-valeur du modèle de Cox univarié ; case entourée de bleu : résultats obtenus dans le cas optimal (*i.e.* C-index médian le plus important) ; case entourée de gris : résultats obtenus sans pré-filtrage.

lasso, elastic net et adaptive elastic net en prenant le C-index comme métrique d'évaluation des capacités de prédiction.

Pour résumer :

— le choix des seuils est variable suivant les cancers.

— le C-index et la p-valeur du modèle de Cox univariée donnent des résultats similaires quant au choix du seuil sur l'EI.

— les pénalisations lasso, elastic net et adaptive elastic net ont des comportements similaires quant au choix du seuil sur l'EI.

— le choix du seuil pour la p-valeur du modèle de Cox univarié est plus instable que celui du seuil sur l'EI suivant les pénalisations.

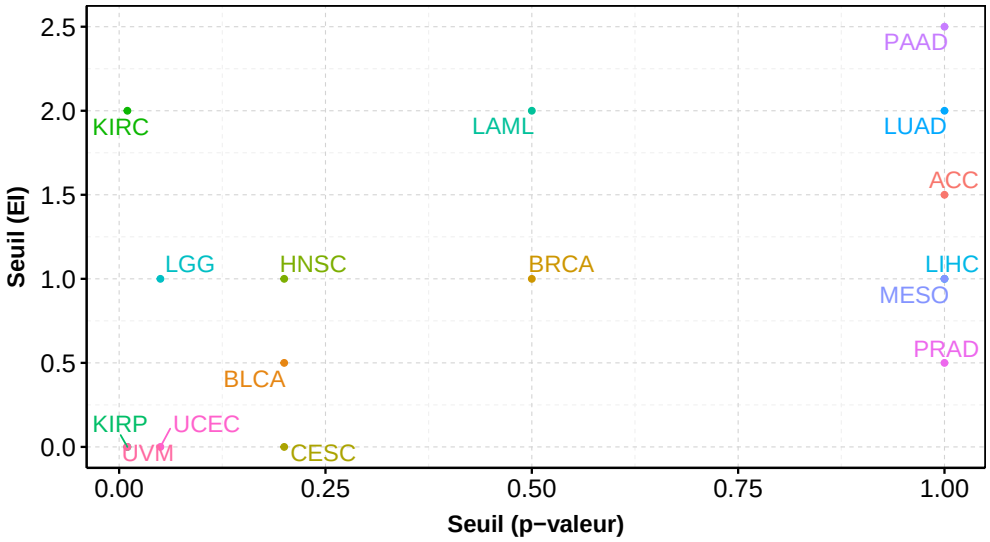


FIGURE 3.3 – **Seuils optimaux pour la pénalisation elastic net, le C-index, et l'ensemble des 16 cancers.**

3.2.2 Impact du pré-filtrage sur les performances de prédiction

Sur les 16 cancers étudiés, le C-index est augmenté de manière significative sur les données de validation pour 11 cancers pour lasso et elastic net, pour 10 cancers pour adaptive elastic net, et pour 4 cancers pour ridge (p-valeur < 0.05, test des rangs signés de Wilcoxon avec correction de Benjamini-Hochberg, Tab. 3.1). Pour BRCA, cette augmentation est significative pour les quatre méthodes de pénalisation, mais pour tous les autres, cette augmentation dépend de la méthode de pénalisation. Par exemple, pour LGG, le C-index augmente significativement pour lasso, elastic net et adaptive elastic net, mais pas pour ridge (Fig. 3.4, Tab. 3.1).

Ensuite, pour les cancers et les méthodes pour lesquels l'augmentation du C-index est significative, le nombre médian de gènes retenu après pré-filtrage est de 914. Le pré-filtrage permet donc d'augmenter les performances de prédiction, mais aussi de réduire l'ensemble des prédicteurs du modèle aux gènes les plus variables et / ou les plus associés à la survie.

Des résultats du même ordre de grandeur sont observés pour la p-valeur du modèle de Cox univarié (diminution significative pour 7 cancers pour ridge et adaptive elastic net, pour 10 cancers pour lasso et elastic net, Tab. Annexe A.5), et pour l'IBS (diminution de l'IBS pour 8 cancers pour elastic net, 5 cancers pour lasso, 11 cancers pour ridge, et 13 cancers pour adaptive elastic net, Tab. Annexe A.6).

Dans la partie suivante, nous nous attacherons à comparer les performances de prédiction des quatre méthodes de pénalisation après pré-filtrage.

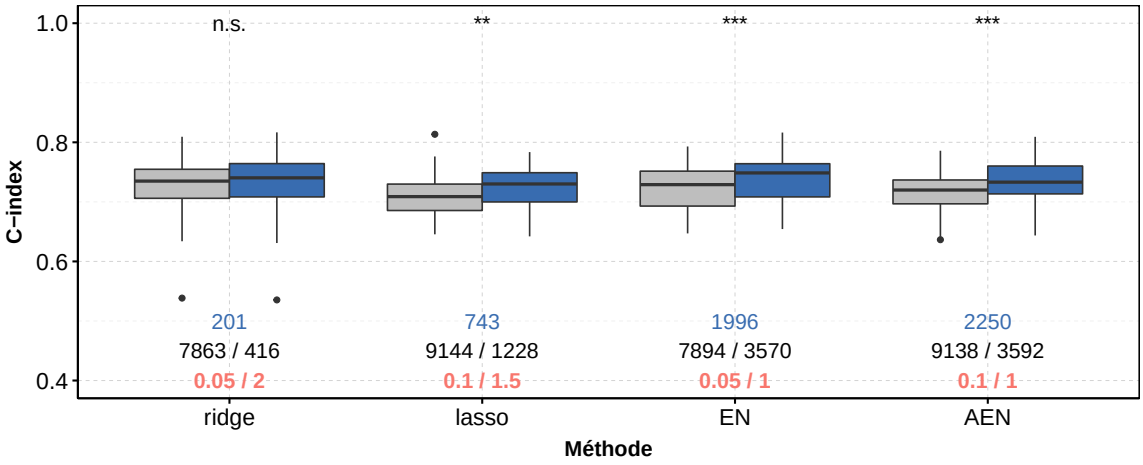


FIGURE 3.4 – **C-index obtenus avant (gris) et après le pré-filtrage maximisant le C-index (bleu) pour les quatre méthodes de pénalisation pour LGG.**

Dans chaque cas, les C-index sont calculés par 10 répétitions d'une validation croisée (K=5), et le pré-filtrage des gènes est fait sur le jeu de données d'apprentissage. Le nombre en bleu en bas du graphique correspond au nombre de gènes retenus après le pré-filtrage dans le cas optimal; les deux nombres en noirs correspondent aux nombres de gènes retenus par la p-valeur du modèle de Cox univarié (gauche), et par le pré-filtrage sur l'EI (droite) dans le cas optimal; en dessous, les deux nombres en rouges correspondent au seuil sur la p-valeur du modèle de Cox univarié (gauche), et sur le pré-filtrage sur l'EI (droite) dans le cas optimal.

Les p-valeurs d'un test des rangs signés de Wilcoxon sont corrigées par la méthode de Benjmaini-Hochberg indépendamment pour chaque méthode suivant les 16 cancers étudiés, et sont indiquées sous forme d'étoiles en haut du graphique.

n.s. : non-significatif; + : p-valeur < 0,1; * : p-valeur < 0,05; ** : p-valeur < 0,01; *** : p-valeur < 0,001.

« EN » : elastic net; « AEN » : adaptive elastic net.

TABLEAU 3.1 – **Augmentation moyenne du C-index après pré-filtrage et niveau de significativité pour les 16 cancers étudiés.**
Les p-valeurs sont calculées par un test des rangs signés de Wilcoxon unilatéral, et corrigées par la méthode de Benjamini-Hochberg pour chacune des méthodes suivant les 16 cancers étudiés.
n.s. : non-significatif; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

Cancer	ridge	lasso	EN	AEN	Cancer	ridge	lasso	EN	AEN
ACC	0	0,0016	0,014	0,023	LGG	0,0038	0,0092	0,017	0,019
	n.s.	n.s.	*	*		n.s.	**	***	***
BLCA	0	0,017	0,022	0,02	LIHC	0,0023	0,02	0,015	0,015
	n.s.	**	***	***		n.s.	**	***	*
BRCA	0,016	0,023	0,01	0,018	LUAD	0,0087	0,013	0,0051	0,019
	***	*	*	***		*	*	n.s.	**
CESC	0,0042	0,033	0	0	MESO	0,0029	0,022	0,0065	0
	n.s.	***	n.s.	n.s.		n.s.	**	n.s.	n.s.
HNSC	0,0061	0,0088	0,01	0,003	PAAD	0,018	0,027	0,02	0,012
	n.s.	+	*	n.s.		*	**	*	n.s.
KIRC	0,0034	0,014	0,0093	0,019	PRAD	0	0,014	0,0031	0,026
	n.s.	**	**	***		n.s.	**	n.s.	***
KIRP	0,0046	0,0018	0,0078	0,0046	UCEC	0,0076	0,004	0	0,01
	n.s.	n.s.	**	**		***	n.s.	n.s.	n.s.
LAML	0,01	0,023	0,02	0,025	UVM	0,014	0	0,012	0
	n.s.	*	*	**		n.s.	n.s.	*	n.s.

3.3 Comparaison des méthodes de pénalisation du modèle de Cox après pré-filtrage

Après le pré-filtrage, les capacités de prédiction des quatre méthodes de pénalisation sont comparables (Fig. 3.5, Fig. Annexe A.10). Notons que la pénalisation ridge obtient des C-index plus faibles que les trois autres méthodes pour KIRP et CESC (p-valeur < 0,01, test de Wilcoxon unilatéral). Pour CESC, lasso et adaptive elastic net obtiennent de meilleures résultats qu'elastic net en prenant le C-index comme métrique d'évaluation des prédictions (p-valeur < 0,01, test de Wilcoxon unilatéral).

Avec l'IBS et la p-valeurs du modèle de Cox univarié, des résultats similaires sont observés : les capacités de prédiction des quatre méthodes de pénalisation sont équivalentes après le pré-filtrage.

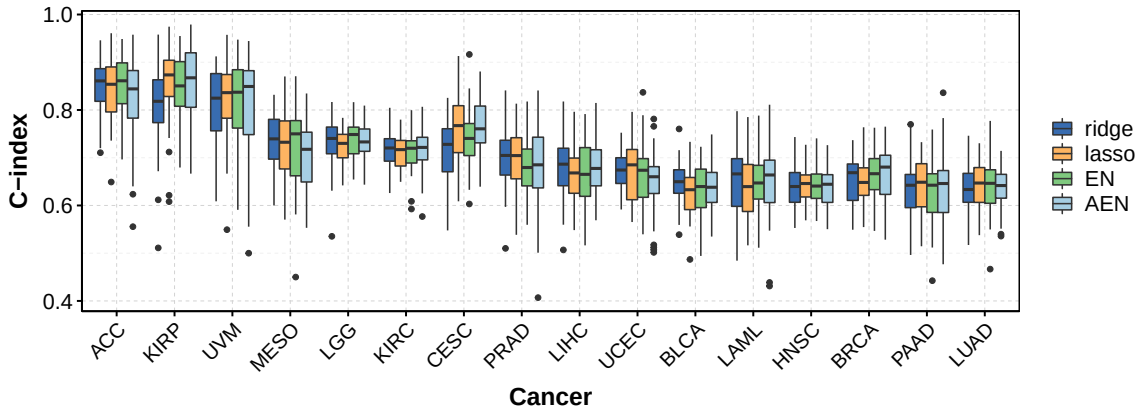


FIGURE 3.5 – **C-index obtenus après prè-filtrage pour les quatre méthodes de pénalisation (*i.e.* ridge, lasso, elastic net et adaptive elastic net) et 16 cancers de TCGA.**
Les cancers sont classés par ordre décroissants du C-index médian calculé avec la pénalisation ridge sans pré-filtrage.

3.4 Comparaison du pré-filtrage bi-dimensionnel à l'algorithme *Iterative Sure Independance Screening*

3.4.1 L'algorithme *Iterative Sure Independance Screening*

L'algorithme *Iterative Sure Independance Screening* (ISIS) est une méthode itérative qui permet de filtrer des gènes et de réduire la dimension [FAN et SONG, 2010]. Cette méthode consiste à classer les gènes suivant leur corrélation individuelle avec la survie (*i.e.* valeurs absolues des coefficients $|\beta_j|$, $j = 1,\ldots,p$, de modèles de Cox univariés), et à ajuster l'ensemble des gènes sélectionnées de manière itérative :

1. calcul des valeurs absolues des coefficients $|\beta_j|$, $j = 1,\ldots,p$, de modèles de Cox univariés et définition d'un ensemble de gènes $A_1 = \{j = 1,\ldots,p / |\beta_j| > \delta_1\}$.

2. utilisation d'un modèle de Cox avec pénalisation lasso sur l'ensemble A_1. Les gènes sélectionnés forment un nouvel ensemble M_1.
3. pour chaque gène j qui n'est pas dans M_1, utilisation d'un modèle de Cox multivarié avec l'ensemble des gènes de M_1 et le gène j. L'apport du gène j qui n'est pas dans M_1 est mesuré par la valeur absolue de son coefficient β_j dans ce modèle de Cox multivarié, et un nouvel ensemble A_2 est définit : $A_2 = \{j \in M_1^c / |\beta_j| > \delta_2\}$, avec M_1^c le complémentaire de M_1. Le nouvel ensemble M_2 est alors définit par l'ensemble des gènes sélectionnés par un modèle de Cox avec pénalisation lasso sur $M_1 \cup A_2$.
4. l'étape 3 est répétée jusqu'à ce que le nombre d'itérations l atteigne un certain niveau maximum l_{max}, ou $M_l = M_{l-1}$ pour $l < l_{max}$, ou $\#M_l > d$, avec « # » la fonction cardinale et d le nombre de gènes désirés à la fin de l'algorithme.

La fonction `SIS` du package `SIS` [SALDANA et FENG, 2018] permet l'utilisation de cette méthode.

Le choix du nombre de gènes d à inclure dans le modèle est basé sur le nombre de patients n du jeu de données d'expression génétique plutôt que sur le modèle utilisé ou la structure des données. Ainsi, comme recommandé par FAN et collab. [2009], nous avons utilisé $d = \lfloor \frac{n}{\log(n)} \rfloor$. Si nous observons une erreur lors de l'appel de la fonction `SIS` avec cette valeur de d, des valeurs de d plus petites sont successivement choisies (*i.e.* $d = \lfloor \frac{n}{2\log(n)} \rfloor$ et $d = \lfloor \frac{n}{4\log(n)} \rfloor$).

Le nombre d'itérations maximales l_{max} est fixé à 10 (défaut) ; le nombre de gènes retenus dans A_1 est fixé à $\lfloor 2d/3 \rfloor$ et permet de définir δ_1 ; le nombre de gènes retenus dans $A_l, l > 1$ est fixé à $d - \#M_{l-1}$ et permet de définir δ_l.

Les deux premières étapes de l'algorithme ci-dessus définissent la méthode *Sure Independance Screening* (SIS), avec δ_1 définit tel que l'ensemble A_1 contiennent d gènes. FAN et SONG [2010] ont montré que sous l'hypothèse d'indépendance des co-variables, la méthode SIS permet de sélectionner un sous-ensemble de gènes contenant la vérité terrain. Les itérations de l'algorithme ISIS permettent de remédier au cas où (i) des gènes fortement corrélés à la survie de manière univarié mais n'apportant pas d'information dans le modèle multivarié existent, ou (ii) des gènes faiblement corrélés à la survie de manière univarié mais important dans le modèle multivarié existent.

Ainsi, nous avons appliqué cet algorithme ISIS sur le jeu de données d'apprentissage pour pré-filtrer les gènes. Nous avons ensuite utilisé un modèle de Cox avec pénalisation ridge sur l'ensemble des gènes restants pour calculer les trois métriques d'évaluation des prédictions (*i.e.* C-index, IBS, p-valeur du modèle de Cox univarié) sur le jeu de données test. Nous avons appliqué cette stratégie à l'ensemble des 16 jeux de données TCGA étudiés dans ce chapitre (partie 3.1.5), et pour chaque cancer, les métriques sont calculées par 10 répétitions d'une validation croisée (K=5) (partie 1.7.2). Nous avons ensuite comparé les résultats obtenus avec ceux présentés dans la partie 3.2 pour le pré-filtrage bi-dimensionnel.

3.4.2 Comparaison avec le pré-filtrage bi-dimensionnel

Tout d'abord, notons que la méthode de filtrage utilisé par l'algorithme ISIS est très proche du filtrage supervisé que nous utilisons. En effet, pour tous les cancers, nous observons une forte corrélation entre les p-valeurs d'un test du rapport de vraisemblance corrigées par la méthode de Benjamini-Hochberg et les valeurs absolues des coefficients $\beta_j, j = 1,\ldots,p$, de modèles de Cox univariés appris individuellement pour chaque gène (Fig. 3.6, données non montrées pour les autres cancers). Les p-valeurs ont l'avantage de prendre en compte l'incertitude (*i.e.* l'écart-type) dans l'estimation des coefficients, mais dépendent de la taille de l'échantillon [ROYALL, 1986; THIESE et collab., 2016].

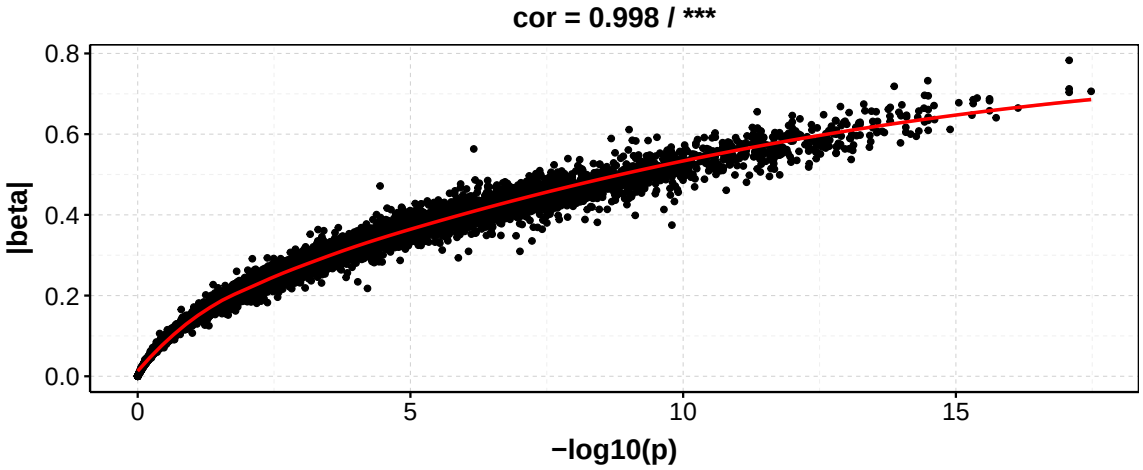


FIGURE 3.6 – **Valeurs absolues des coefficients β en fonction des p-valeurs en échelle $-\log_{10}$ d'un test du rapport de vraisemblance corrigées par la méthode de Benjamini-Hochberg calculés à partir d'un modèle de Cox univarié pour LGG.**
La courbe *loess* est tracée en rouge, la corrélation de Spearman et le niveau de significativité de cette corrélation sont notés dans le titre du graphique (*** : p-valeur < 0,001).

Ensuite, le nombre médian de gènes rejetés par la méthode ISIS est beaucoup plus important que celui obtenu après le pré-filtrage bi-dimensionnel (Fig. 3.7 et Fig. Annexe A.12). En moyenne sur l'ensemble des cancers et des quatre méthodes de pénalisation (*i.e.* ridge, lasso, elasti net, adaptive elastic net), le nombre de gènes retenus est de 914 pour le filtrage bi-dimensionnel, et de 28 pour ISIS. L'objectif de l'algorithme ISIS est de pré-filtrer les gènes pour réduire la dimension et obtenir un sous-ensemble d'au plus $d = \lfloor \frac{n}{\log(n)} \rfloor$ covariables inférieure au nombre d'échantillons n. Par exemple, pour une cohorte de taille $n = 200$, on obtient $d = 37$. Si un groupe de plus de 37 gènes fortement corrélés entre eux et à la survie existe dans le jeu de données, ces gènes vont tous être sélectionnés par le pré-filtrage. L'information contenue dans les autres gènes moins corrélés à la survie de manière univarié, mais tout de même intéressants dans le modèle multivarié, sera perdue. Le processus itératif de l'algorithme ISIS permet de remédier en partie à ce biais, mais les gènes sont ajoutés un par un au modèle multivarié sur au plus 10 itérations, et

l'information contenue dans les différents groupes de gènes se trouve « noyée » et n'est pas totalement prise en compte dans le modèle multivarié.

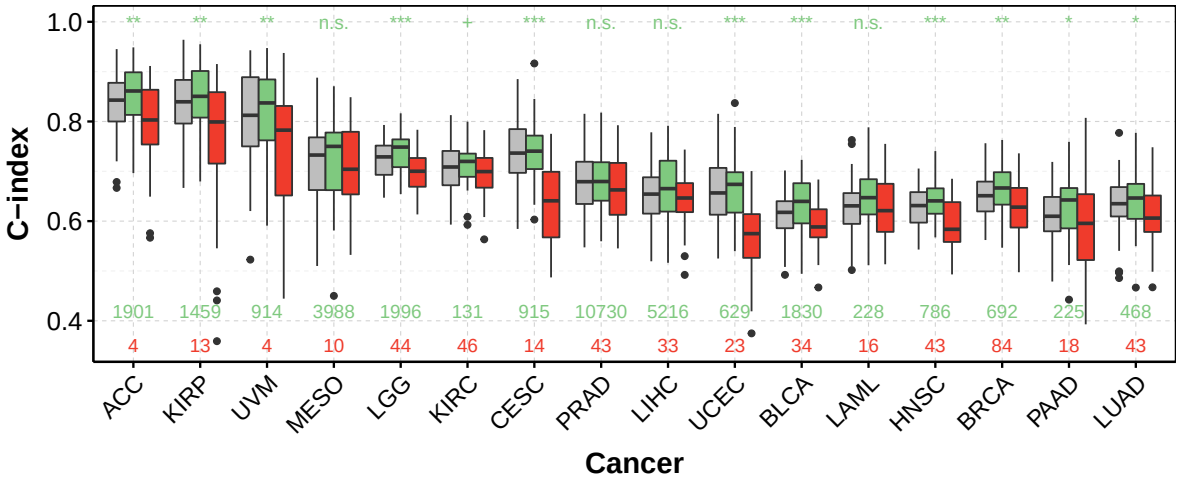


FIGURE 3.7 – **C-index obtenus sans pré-filtrage (gris), avec le pré-filtrage bi-dimensionnel et la pénalisation elastic net (vert), et avec l'algorithme *Sure Independance Screening* (rouge) pour les 16 cancers de TCGA étudiés.**

Dans chaque cas, les C-index sont calculés par 10 répétitions d'une validation croisée (K=5). Les nombres rouges et verts indiquent le nombre de gènes après filtrage par la méthode ISIS et bi-dimensionnel, respectivement. Les p-valeurs corrigées par la méthode de Benjamini-Hochberg permettant de tester si les C-index médians obtenus avec les deux méthodes sont significativement différents (vert : le C-index médian obtenu avec le filtrage bi-dimensionnel est le plus élevé, rouge : le C-index médian obtenu avec la méthode ISIS est plus élevé).
n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

De plus, le nombre d de gènes retenus dans le modèle dépend du nombre de patients, et ne prend pas en compte la diversité qui existe entre les jeux de données. Pour certains cancers, très peu de gènes peuvent être associés à la survie, alors que pour d'autres, un nombre important de gènes peuvent apporter de l'information et un pouvoir prédictif. Dans notre algorithme, le seuil, et donc le nombre de gènes retenus par le pré-filtrage, est fixé par validation croisé afin d'optimiser les prédictions. Ce seuil va donc dépendre du jeu de données utilisé, et seuls les gènes inutiles pour optimiser les prédictions vont être filtrés. Le nombre de gènes restants est donc plus important avec le pré-filtrage bi-dimensionnel (914 en moyenne) qu'avec l'algorithme ISIS (28 en moyenne). L'utilisation d'une pénalisation lasso, elastic net ou adaptive elastic net permet ensuite de réduire l'ensemble des prédicteurs à une cinquantaine de gènes.

Ainsi, les performances de prédiction obtenues avec le pré-filtrage bi-dimensionnel sont meilleures que celles obtenues avec le pré-filtrage par l'algorithme ISIS (Fig. 3.7 et Fig. Annexe A.12). Le C-index médian obtenu après pré-filtrage bi-dimensionnel et elastic net est significativement supérieur pour 11 cancers (ACC, KIRP, UVM, LGG, CESC,

UCEC, BLCA, HNSC, BRCA, PAAD, LUAD, p-valeur < 0.05 après correction de Benjamini-Hochberg, test de Wilcoxon), et la p-valeur médiane du modèle de Cox univarié est significativement plus petite pour 11 cancers (ACC, KIRP, UVM, LGG, CESC, UCEC, BLCA, LAML, HNSC, BRCA, LUAD, p-valeur < 0.05 après correction de Benjamini-Hochberg, test de Wilcoxon). Pour l'IBS, les résultats sont équivalents pour l'ensemble des cancers, à l'exception de CESC pour lequel l'IBS est significativement plus petit avec le pré-filtrage bi-dimensionnel.

Finalement, le pré-filtrage bi-dimensionnel, basé à la fois sur la corrélation individuelle avec la survie et la variabilité entre les patients, permet de fixer les seuils de filtrage par validation croisée pour optimiser les prédictions. De plus, cette approche permet de réduire l'ensemble des co-variables à un sous-ensemble de gènes pertinents pour prédire la survie, sans perdre l'information contenue dans les interactions entre les différents groupes de gènes et prise en compte dans le modèle multivarié.

3.5 Conclusions et perspectives

Calculer les seuils de pré-filtrage par validation croisée sur l'EI après stabilisation de la variance et sur les p-valeurs du modèle de Cox univarié après correction par Benjamini-Hochberg permet :

— de réduire simplement le nombre de prédicteurs à inclure dans le modèle de Cox pénalisé.

— de ne garder que les gènes les plus pertinents pour prédire la survie.

— d'augmenter la stabilité des gènes qui composent les scores de risque (indices pronostiques).

— d'optimiser les capacités de prédiction.

Nous avons vu que le pré-filtrage permet d'améliorer les prédictions pour environ la moitié des cancers étudiés pour chacune des méthodes de pénalisation (*i.e.* ridge, lasso, elastic net et adaptive elastic net, test des rangs signés de Wilcoxon) et en prenant le C-index ou la p-valeur du modèle de Cox univarié comme métrique d'évaluation. Cependant, cette augmentation reste marginale en valeur absolue. En effet, l'augmentation moyenne du C-index pour l'ensemble des 16 cancers étudiés dans ce chapitre n'est que de 0,016 pour la pénalisation lasso.

Ensuite, l'algorithme SIS (sans itération) est équivalent au pré-filtrage supervisé sur les p-valeurs corrigées du modèle de Cox univarié, mais sans calibration du seuil par validation croisée. L'algorithme ISIS (avec itérations) permet de remédier au cas où (i) des gènes fortement corrélés à la survie de manière univarié mais n'apportant pas d'information dans le modèle multivarié existent, ou (ii) des gènes faiblement corrélés à la survie de manière univarié mais important dans le modèle multivarié existent. Cependant, le pro-

cessus itératif rend difficile en pratique la calibration des seuils de pré-filtrage, et l'heuristique choisi ne prend pas en compte la spécificité du jeu de données étudié. Les performances de prédiction obtenues avec l'algorithme ISIS sont moins bonnes que celles obtenues avec le pré-filtrage bi-dimensionnel. Le nombre de gènes retenus dans le modèle multivarié semble trop faible. De plus, le pré-filtrage sur l'EI permet, pour certains cancers, d'obtenir de meilleures prédictions et demeure une composante importante du filtrage bi-dimensionnel.

Dans notre étude, nous avons retenu la corrélation avec la survie et l'EI comme critères de pré-filtrage. D'autres métriques peuvent être retenues suivant les besoins, le type de données, et les applications. Par exemple, le C-index calculé de manière univarié pour chaque gène sur le jeu de données d'apprentissage peut être intéressant à considérer comme score de pré-filtrage. De plus, le niveau médian d'expression peut être intéressant à considérer. En effet, les gènes fortement exprimés peuvent être séquencés avec une profondeur de séquençage faible, et les coûts peuvent être réduits [MILANEZ-ALMEIDA et collab., 2020]. L'étude de l'impact de la profondeur de séquençage sur les capacités de prédiction du modèle de Cox fera l'objet du prochain chapitre.

Enfin, nous calculons les seuils de pré-filtrage grâce aux médianes des C-index (reps. des p-valeurs du modèle de Cox univarié, des IBS) obtenu(e)s. Cependant, il serait intéressant de prendre en considération la variance de ces métriques dans le score permettant d'évaluer la qualité des prédictions et de fixer les seuils. En effet, un modèle qui obtient un C-index médian élevé mais avec une variance importante est peu utile en pratique.

3.6 Références

ANDERS, S. et W. HUBER. 2010, «Differential expression analysis for sequence count data», *Genome Biology*, vol. 11, nº 10, doi :10.1186/gb-2010-11-10-r106, p. R106, ISSN 1474-760X. URL https://doi.org/10.1186/gb-2010-11-10-r106. 94

BENJAMINI, Y. et Y. HOCHBERG. 1995, «Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, nº 1, p. 289–300, ISSN 0035-9246. URL https://www.jstor.org/stable/2346101, publisher : [Royal Statistical Society, Wiley]. 93, 96

BENNER, A. et collab.. 2010, «High-dimensional Cox models : the choice of penalty as part of the model building process», *Biometrical Journal. Biometrische Zeitschrift*, vol. 52, nº 1, doi :10.1002/bimj.200900064, p. 50–69, ISSN 1521-4036. 92

BOULESTEIX, A.-L. et collab.. 2020, «Statistical learning approaches in the genetic epidemiology of complex diseases», *Human Genetics*, vol. 139, nº 1, doi :10.1007/

s00439-019-01996-9, p. 73–84, ISSN 1432-1203. URL https://doi.org/10.1007/s00439-019-01996-9. 92, 96

BØVELSTAD, H. M. et collab.. 2007, «Predicting survival from microarray data–a comparative study», *Bioinformatics (Oxford, England)*, vol. 23, nº 16, doi :10.1093/bioinformatics/btm305, p. 2080–2087, ISSN 1367-4811. 92

DAI, L. et collab.. 2017, «MiR-221, a potential prognostic biomarker for recurrence in papillary thyroid cancer», *World Journal of Surgical Oncology*, vol. 15, doi :10.1186/s12957-016-1086-z, ISSN 1477-7819. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5219708/. 93

DOMANY, E. 2014, «Using High-Throughput Transcriptomic Data for Prognosis : A Critical Overview and Perspectives», *Cancer Research*, vol. 74, nº 17, doi :10.1158/0008-5472.CAN-13-3338, p. 4612–4621, ISSN 0008-5472, 1538-7445. URL http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-13-3338. 92

FA, B. et collab.. 2019, «Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma», *EBioMedicine*, vol. 44, doi :10.1016/j.ebiom.2019.05.010, p. 250–260, ISSN 2352-3964. URL http://www.sciencedirect.com/science/article/pii/S2352396419303111. 93, 95

FAN, J. et R. SONG. 2010, «Sure independence screening in generalized linear models with NP-dimensionality», *Annals of Statistics*, vol. 38, nº 6, doi :10.1214/10-AOS798, p. 3567–3604, ISSN 0090-5364, 2168-8966. URL https://projecteuclid.org/euclid.aos/1291126966, publisher : Institute of Mathematical Statistics. 102, 103

FAN, J. et collab.. 2009, «Ultrahigh Dimensional Feature Selection : Beyond The Linear Model», *The Journal of Machine Learning Research*, vol. 10, p. 2013–2038, ISSN 1532-4435. 103

JIANG, Y. et collab.. 2016, «Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis», *Genomics*, vol. 107, nº 6, doi :10.1016/j.ygeno.2016.04.005, p. 223–230, ISSN 1089-8646. 93, 95

LIAO, Q. et collab.. 2011, «Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network», *Nucleic Acids Research*, vol. 39, nº 9, doi :10.1093/nar/gkq1348, p. 3864–3878, ISSN 1362-4962. 93, 95

LOVE, M. I. et collab.. 2014, «Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2», *Genome Biology*, vol. 15, nº 12, doi :10.1186/s13059-014-0550-8, p. 550, ISSN 1474-760X. URL https://doi.org/10.1186/s13059-014-0550-8. 95

MICHIELS, S. et collab.. 2005, «Prediction of cancer outcome with microarrays : a multiple random validation strategy», *Lancet (London, England)*, vol. 365, nº 9458, doi :10.1016/S0140-6736(05)17866-0, p. 488–492, ISSN 1474-547X. 93, 95

MILANEZ-ALMEIDA, P. et collab.. 2020, «Cancer prognosis with shallow tumor RNA sequencing», *Nature Medicine*, vol. 26, nº 2, doi :10.1038/s41591-019-0729-3, p. 188–192, ISSN 1078-8956, 1546-170X. URL http://www.nature.com/articles/s41591-019-0729-3. 107

MILLER, L. D. et collab.. 2002, «Optimal gene expression analysis by microarrays», *Cancer Cell*, vol. 2, nº 5, doi :10.1016/S1535-6108(02)00181-2, p. 353–361, ISSN 1535-6108. URL http://www.sciencedirect.com/science/article/pii/S1535610802001812. 92, 93

RAMAN, P. et collab.. 2019, «A comparison of survival analysis methods for cancer gene expression RNA-Sequencing data», *Cancer Genetics*, vol. 235-236, doi :10.1016/j.cancergen.2019.04.004, p. 1–12, ISSN 2210-7762. URL http://www.sciencedirect.com/science/article/pii/S2210776218304897. 93

ROYALL, R. M. 1986, «The Effect of Sample Size on the Meaning of Significance Tests», *The American Statistician*, vol. 40, nº 4, doi :10.1080/00031305.1986.10475424, p. 313–315, ISSN 0003-1305. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1986.10475424, publisher : Taylor & Francis _eprint : https ://www.tandfonline.com/doi/pdf/10.1080/00031305.1986.10475424. 104

SALDANA, D. F. et Y. FENG. 2018, «SIS : An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models», *Journal of Statistical Software*, vol. 83, nº 1, doi :10.18637/jss.v083.i02, p. 1–25, ISSN 1548-7660. URL https://www.jstatsoft.org/index.php/jss/article/view/v083i02, number : 1. 103

SHAO, Y. et collab.. 2019, «Serum miR-22 Could be a Potential Biomarker for the Prognosis of Breast Cancer», *Clinical Laboratory*, vol. 65, nº 4, doi :10.7754/Clin.Lab.2018.180825, ISSN 1433-6510. 93

THIESE, M. S. et collab.. 2016, «P value interpretations and considerations», *Journal of Thoracic Disease*, vol. 8, nº 9, doi :10.21037/jtd.2016.08.16, p. E928–E931, ISSN 2072-1439. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059270/. 104

ZHAO, Q., X. SHI, Y. XIE, J. HUANG, B. SHIA et S. MA. 2015, «Combining multidimensional genomic measurements for predicting cancer prognosis : observations from TCGA», *Briefings in Bioinformatics*, vol. 16, nº 2, doi :10.1093/bib/bbu003, p. 291–303, ISSN 1467-5463. URL https://academic.oup.com/bib/article/16/2/291/246070, publisher : Oxford Academic. 93, 95

ZWIENER, I., B. FRISCH et H. BINDER. 2014, «Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures», *PLoS ONE*, vol. 9, nº 1, doi :10.1371/journal.pone.0085150, p. e85 150, ISSN 1932-6203. URL https://dx.plos.org/10.1371/journal.pone.0085150. 94

Chapitre 4

Impact de la profondeur de séquençage des données miRNA-seq sur la prédiction

Sommaire

4.1 Contexte de l'étude . . . 112
4.1.1 Intérêt des miARN dans le cancer . . . 112
4.1.2 La taille des banques . . . 112
4.1.3 Calibration des études transcriptomiques . . . 113
4.1.4 Objectifs du chapitre . . . 114
4.2 Choix des cancers étudiés . . . 114
4.2.1 Critères de qualité des données pour prédire la survie . . . 114
4.2.2 Variabilité de la taille des banques . . . 116
4.3 Comparaison des prédictions obtenues avec les ARNm et les miARN . . . 118
4.4 Prédiction de la survie avec les variables cliniques et les données miRNA-seq . . . 121
4.5 Dégradation des données de séquençage . . . 123
4.5.1 Dégradation de la profondeur de séquençage . . . 123
4.5.2 Diminution du nombre de patients . . . 124
4.6 Impact de la taille des banques et du nombre de patients sur la prédiction 125
4.6.1 Impact de la taille des banques . . . 126
4.6.2 Impact du nombre de patients . . . 129
4.7 Cause de la dégradation de la qualité de prédiction . . . 129
4.7.1 Deux hypothèses envisagées . . . 129
4.7.2 Test de la première hypothèse . . . 130
4.7.3 Test de la deuxième hypothèse . . . 131
4.8 Conclusions . . . 133
4.9 Références . . . 134

4.1 Contexte de l'étude

4.1.1 Intérêt des miARN dans le cancer

Nous avons décri les miARN dans l'introduction (1.2.3). Pour rappel, ce sont de petites molécules d'ARN (20 à 24 nucléotides) non codantes qui interviennent principalement dans la régulation post-transcriptionelles des gènes (*i.e.* les miARN agissent directement sur les molécules d'ARNm) [BARTEL, 2018].

CALIN et CROCE [2006] ont mis en évidence une association entre le niveau d'expression des miARN et la survie. Depuis, les miARN ont étaient largement étudiés, et apparaissent comme une cible thérapeutique intéressante dans le cancer [RUPAIMOOLE et SLACK, 2017]. En effet, de nombreuses études ont utilisé les niveaux d'expression des miARN dans le modèle de Cox pour établir des signatures de gènes prédictives de la survie [SHI et collab., 2018; YU et collab., 2008; ZHANG et collab., 2015; ZHOU et collab., 2016]. Enfin, l'intérêt pour les miARN dans les études portant sur le cancer ne cesse de croître depuis 2005 : le nombre de papiers référencés dans PubMed contenant les mot « *miRNA* » et « *cancer* » dans leur titre ou leur résumé est passé de 20 en 2005 à 2734 en 2019 (Fig. 4.1).

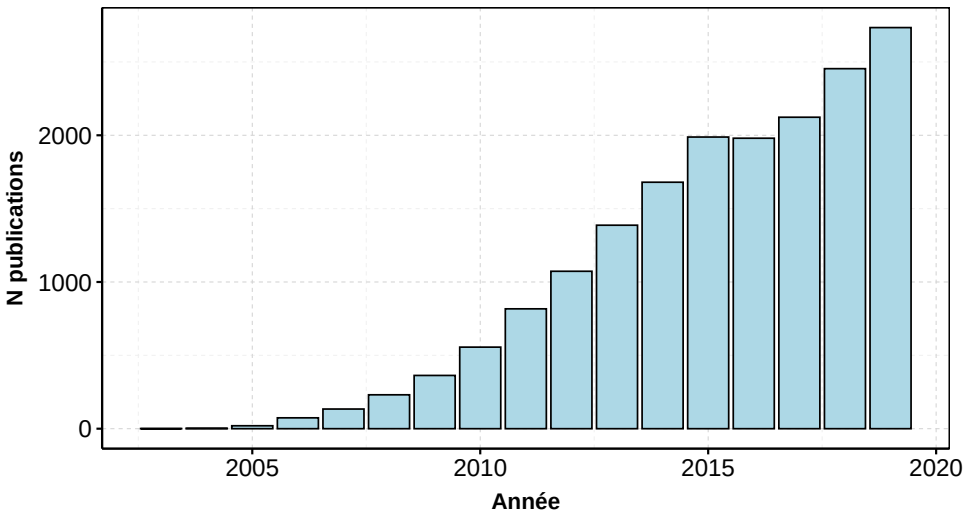


FIGURE 4.1 – **Nombre de publications référencées dans PubMed et publiées entre 2003 et 2019 contenant les mots-clefs « *miRNA* » et « *cancer* » dans leur titre ou leur abstract.**
La recherche associée au nombre de publications est "(miRNA[Title/Abstract]) AND (cancer[Title/Abstract])" dans le moteur de recherche de PubMed, et a était faite le 18/08/2020.

4.1.2 La taille des banques

Nous avons fait une description technique du séquençage à haut débit en introduction (partie 1.2). Nous rappelons ici que pour un patient donné, la taille de la banque correspond au nombre total de lectures alignées sur un génome de référence. Elle peut

varier suivant les expériences (et donc ici les patients), et est équivalente à la notion de « profondeur de séquençage ». Ces deux nomenclatures seront utilisées de manière interchangeable dans la suite du texte.

4.1.3 Calibration des études transcriptomiques

Les deux paramètres essentiels à calibrer lors de l'élaboration d'une étude basée sur les transcriptomes de biopsies d'une cohorte de patients sont la taille des banques et le nombre de patients, et cette calibration s'effectue sous contrainte de coûts.

La taille des banques est un paramètre central du séquençage haut débit : plus cette taille est importante, plus le nombre de gènes détectés sera important et plus précises seront les mesures du niveau d'expression, mais plus le coût sera élevé [Mortazavi et collab., 2008; Sims et collab., 2014]. Bien que les coûts de la technologie RNA-seq aient diminué drastiquement au cours de la dernière décennie, ils restent trop élevés pour que le séquençage soit utilisé de manière routinière en clinique [Cieślik et Chinnaiyan, 2018; Kumar-Sinha et Chinnaiyan, 2018; Michiels et collab., 2016; Senft et collab., 2017].

Pour un budget donné, connaître la taille de banque minimale permettant d'obtenir les performances nécessaires permet de maximiser le nombre de patients de la cohorte. Plus le nombre de patients est élevé, plus la puissance statistique est importante, et plus les prédictions obtenues seront réalistes. Les grandes cohortes de patients permettent aussi d'utiliser des algorithmes modernes tel que les réseaux de neurones profonds, de prendre en compte des termes d'interaction avec des variables cliniques ou des traitements [Kurtz et collab., 2019], ou encore de stratifier les patients suivant des caractéristiques cliniques (âge, sexe, traitements reçus, etc.) ou transcriptomiques. D'autres applications peuvent être envisagées, comme le séquençage longitudinal (*i.e.* à différents instants t) des patients [McGranahan et Swanton, 2017], et / ou le séquençage de différentes zones de la tumeur [Andor et collab., 2016; Gerlinger et collab., 2012] (hétérogénéité intra-tumorale). Pour résumé, il y a donc un compromis à trouver entre la profondeur du séquençage et le nombre d'échantillons séquencés.

Ensuite, lorsque des données sont analysées, une appréciation de l'impact de la taille des banques sur les métriques d'évaluation choisies à travers les « courbes de saturation » [Tarazona et collab., 2011] permet de vérifier que les résultats ne sont pas sous-optimaux du fait d'une profondeur de séquençage trop faible. Par exemple, Bass et collab. [2019] ont montré que le nombre de gènes différentiellement exprimés n'était pas saturé en fonction de la taille des banques dans certaines études. Plus de gènes auraient été détectés avec une profondeur de séquençage plus importante, et de l'information est donc perdue à cause d'une profondeur de séquençage trop faible des données. D'autres chercheurs se sont intéressés à l'impact de la profondeur de séquençage [Liu et collab., 2014; Rapaport et collab., 2013; Tarazona et collab., 2011] et du nombre de patients [Pawitan et collab., 2005] sur la détection de gènes différentiellement exprimés et de voies de signalisation

biologiquement pertinentes [HEIMBERG et collab., 2016; KLIEBENSTEIN, 2012].

Enfin, MILANEZ-ALMEIDA et collab. [2020] ont montré que séquencer quelques centaines de milliers de lectures par échantillon suffit à converger vers les prédictions optimales avec un modèle de Cox et une pénalisation elastic net. Cela revient à diviser la taille des banques des données mRNA-seq de TCGA d'un facteur 100, ce qui correspond en moyenne à 500 000 lectures.

4.1.4 Objectifs du chapitre

Ainsi, plusieurs questions émergent concernant la calibration d'une étude clinique qui repose sur le séquençage haut débit des miARN :

— quelles capacités de prédiction peut-on attendre pour une taille de banque et un nombre de patients donnés ?

— pour un budget donné, quel est le meilleur compromis entre la taille des banques et le nombre de patients permettant de maximiser les capacités de prédiction ?

Dans ce contexte, l'objectif de ce chapitre est d'étudier l'impact du nombre de patients et de la taille des banques des données RNA-seq de miARN sur les capacités de prédiction du modèle de Cox pénalisé. En d'autres termes, le but est d'optimiser les prédictions et le nombre d'échantillons séquencés sous contrainte de coûts..

Dans la suite de ce chapitre, nous utiliserons la pénalisation ridge (VERWEIJ et VAN HOUWELINGEN [1994], partie 1.8.3). Si un modèle parcimonieux est souhaité, la même démarche peut être adoptée avec la pénalisation elastic net. Nous avons remarqué au chapitre 2 que les prédictions obtenues avec elastic net et ridge sont très similaires. Les conclusions que nous allons tirer de notre analyse devraient donc être très similaire avec la pénalisation elastic net. Plus généralement, la méthodologie que nous allons présenter peut s'étendre à tout type d'algorithme de prédiction de la survie.

Les données utilisées proviennent de la base de données TCGA (`https://www.cancer.gov/tcga`). Pour le séquençage des miARN et afin de comparer différentes études portant sur le séquençage des miARN, CHU et collab. [2016] ont détaillés en détail les technologies et la procédure RNA-seq mise en place.

4.2 Choix des cancers étudiés

4.2.1 Critères de qualité des données pour prédire la survie

Pour les données miRNA-seq de la base de données TCGA, nous avons retenu 25 cancers suivant les critères décrits en introduction (partie 1.5.2). Nous utilisons ensuite la même méthodologie qu'au chapitre précédent (partie 3.1.5 pour ne retenir qu'un nombre restreint de cancers).

Ensuite, nous avons choisi de n'utiliser que les cancers pour lesquels les données transcriptomiques de miARN permettent de prédire correctement la survie dans le modèle de Cox avec pénalisation ridge. En effet, ce sont pour ces cancers que les données du niveau d'expression des miARN ont le plus d'applications et de valorisations potentielles. Dans ce sens, nous considérons qu'un jeu de données a un pouvoir prédictif si le C-index médian est significativement supérieur à 0,6. Ainsi, nous avons calculé 50 C-index par 10 répétitions d'une validation croisée (K=5) pour chaque cancer avec les données miRNA-seq (Fig. 1.6). Un test de Wilcoxon unilatéral nous permet alors de déterminer si la médiane m des C-index obtenus est significativement supérieure à 0,6. Ce test a pour hypothèse nulle $H_0 : m < 0,6$, et pour hypothèse alternative $H_1 : m > 0.6$. La méthode de Benjamini-Hochberg (partie 1.5.3, BENJAMINI et HOCHBERG [1995]) nous permet de corriger les p-valeurs obtenues pour les 25 cancers.

Suite à ces tests et au niveau 5%, nous avons sélectionné 11 cancers parmi les 25 (Fig. 4.2). Pour s'assurer d'une bonne association des données avec la survie, nous avons comparé les distributions des C-index obtenues avec et sans permutations par un test de Wilcoxon unilatéral. Ce test permet de vérifier que la médiane des C-index obtenus est significativement supérieure à la médiane des C-index obtenus s'il n'y avait aucun lien entre les données transcriptomiques et la survie. Permuter les patients dans les données de survie sans toucher aux données de transcriptomique permet en effet de décorréler ces deux types de données. Les p-valeurs corrigées par la procédure de Benjmaini-Hochberg pour les 25 cancers restent significatives pour les 11 cancers retenus (données non montrées).

Les caractéristiques de ces 11 cancers sont reportées dans le Tableau 4.1.

TABLEAU 4.1 – **Caractéristiques des onze cancers étudiés dans le chapitre 4.**
Nous calculons les C-index médians par 10 répétitions d'une validation croisée (K=5) avec l'ensemble des gènes (miARN) et le modèle de Cox avec pénalisation ridge. Les cancers sont classés par ordre décroissant de ces C-index médians.

Cancer	C-index estimé	n (#patients)	p (#miARN)	Taux de censure	Taux de survie à 3 ans
ACC	0,83	79	499	0,65	0,76
UVM	0,82	80	502	0,71	0,72
KIRP	0,79	287	461	0,85	0,87
MESO	0,72	85	495	0,14	0,19
KIRC	0,70	513	462	0,67	0,76
LGG	0,70	507	500	0,62	0,56
CESC	0,69	291	526	0,76	0,72
LIHC	0,67	361	529	0,65	0,63
PRAD	0,67	493	452	0,81	0,80
UCEC	0,65	534	553	0,84	0,83
BLCA	0,64	403	561	0,56	0,49

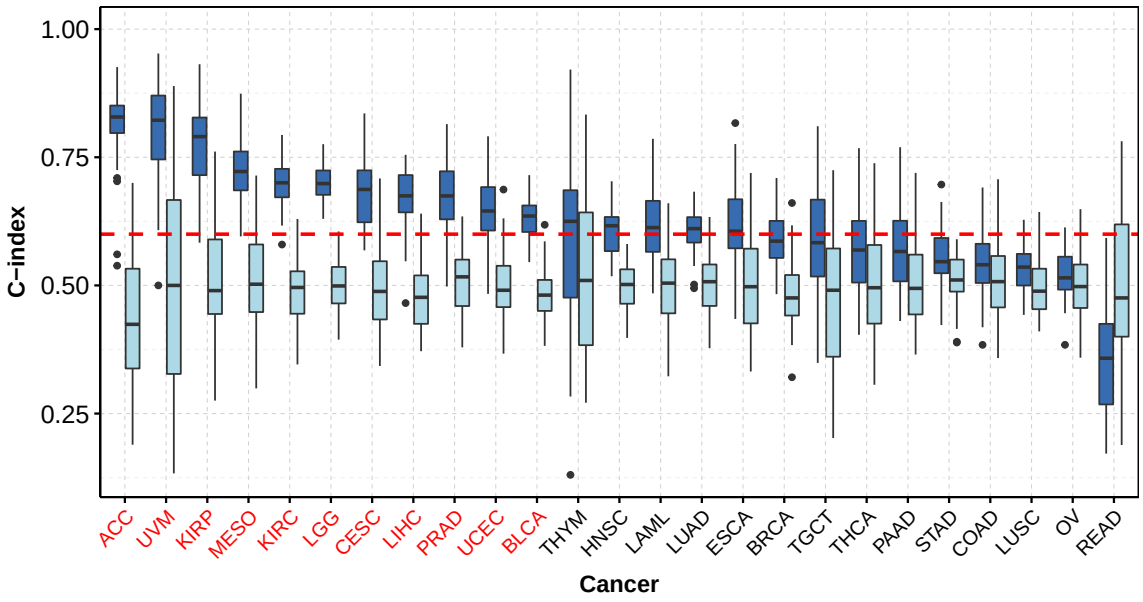


FIGURE 4.2 – ***Boxplot*** **des C-index pour 25 cancers de TCGA obtenus avec les données miRNA-seq.**
Les C-index sont calculés sans permutations (bleu foncé) et avec permutations (bleu clair). Chaque boite contient 50 C-index calculés par 10 répétitions d'une validation croisée (K=5) avec l'ensemble des gènes (miARN) et le modèle de Cox avec pénalisation ridge. Les noms des 11 cancers écrits en rouge sont ceux pour lesquels la p-valeur corrigée par la procédure de Benjmaini-Hochberg pour l'ensemble des 25 tests de wilcoxon unilatéraux permettant de comparer le C-index médian à 0,6 (ligne pointillée horizontale rouge) est inférieure au niveau choisi $\alpha = 0,05$.

4.2.2 Variabilité de la taille des banques

Tout d'abord, la taille des banques varie suivant les expériences (et donc les patients) pour un cancer donné (Fig. 4.3.A et Fig. Annexe A.13). Comme nous l'avons mentionné en introduction (partie 1.3.7), ces différences peuvent s'expliquer par certains biais techniques du séquençage haut-débit :

— quelques fragments d'ADN ne reçoivent pas d'adaptateur, et ne vont pas être séquencés.

— certaines lectures ne sont pas alignées sur le génome (*e.g.* problème de séquençage de l'adaptateur qui empêche la lecture d'être associée à un patient), et leur nombre peut varier d'un patient à l'autre.

— l'amplification PCR de séquences d'ADN riches en bases nucléotidiques guanine (G) et cytosine (C) est difficile [MAMMEDOV et collab., 2008], et implique un biais dans les données RNA-seq [BENJAMINI et SPEED, 2012; RISSO et collab., 2011].

— le nombre de fragments peut varier d'une grille à l'autre, et le multiplexage peut induire des nombre de lectures différentes pour deux patients dont les échantillons ne sont pas sur la même grille.

Ensuite, la taille des banques est équivalente suivant les cancers, à quelques exceptions près, et se répartie autour de 5×10^6 pour les miARN (Fig. 4.3.A), et de 5×10^7 pour les ARNm (Fig. Annexe A.13). La taille des banques des ARNm est donc supérieure à celle des miARN d'un facteur 10 en moyenne, et il n'y a pas de lien particulier entre la profondeur de séquençage choisie pour les ARNm et les miARN entre les différents cancers. Notons que pour LAML, nous observons une profondeur de séquençage plus faible que pour les autres cancers (Fig. 4.3.B). De plus, les données d'ARNm et de miARN se compose respectivement de 20 000 et 500 gènes en moyenne. Ainsi, il y a en moyenne 4 fois plus de lectures alignées par gène pour les miARN que pour les ARNm.

Enfin, quelques différences notables sont à noter concernant la taille des banques entre cancers. Par exemple, pour les miARN, la taille médiane des banques est de 720 000 lectures pour LAML, 2,5 millions pour KIRC et 7,5 millions pour LGG (Fig. 4.3).

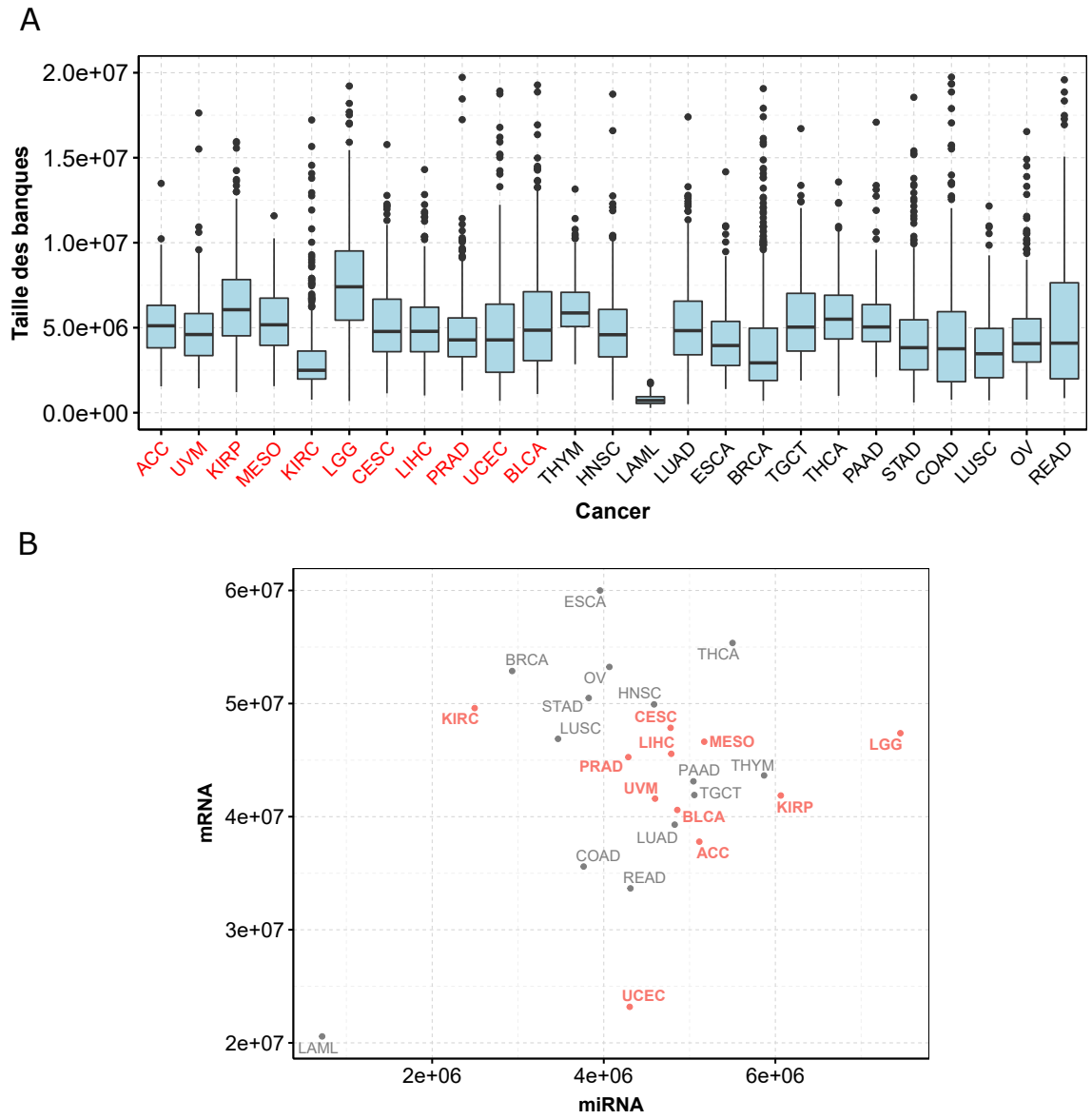


FIGURE 4.3 – **Tailles des banques pour 25 cancers de TCGA obtenues avec les données RNA-seq (miARN et mARN).**
Les 11 cancers dont le nom est écrit en rouge sont ceux qui seront étudiés par la suite. (A) *Boxplot* des tailles des banques pour 25 cancers de TCGA, obtenu avec les données miRNA-seq. Les cancers sont classés par ordre décroissant du C-index médian obtenu après un modèle de Cox avec pénalisation ridge sur les données miARN. (B) Médianes des tailles des banques pour 25 cancers de TCGA obtenues avec les données mRNA-seq (ordonnée) et miRNA-seq (abscisse).

4.3 Comparaison des prédictions obtenues avec les ARNm et les miARN

Les 25 C-index médians obtenus avec les données RNA-seq pour les ARNm et les miARN sont fortement corrélés (corrélation de Pearson de 0,95 - p-valeur = $2,7 \times 10^{-13}$, test de corrélation de Pearson, Fig. 4.4.A). Les différentes capacités de prédiction proviennent ainsi plus du choix du cancer que du type de données utilisées (*i.e.* miARN ou ARNm).

Pour approfondir ce premier résultat, nous avons utilisé un test de Wilcoxon permettant de comparer les C-index médians obtenus avec les données RNA-seq de miARN et d'ARNm pour l'ensemble des 25 cancers (Fig. 4.4.B). Nous avons effectué une correction de tests multiples par la méthode de Benjamini-Hochberg [BENJAMINI et HOCHBERG, 1995] pour l'ensemble de ces 25 tests. Ainsi, avec vingt fois moins de prédicteurs en moyenne, les miARN permettent de prédire la survie de manière équivalentes aux ARNm pour 16 cancers sur 25 (ACC, UVM, MESO, KIRC, CESC, LIHC, PRAD, UCEC, BLCA, THYM, HNSC, LAML, LUAD, TGCT, LUSC, READ). Pour 8 cancers (KIRP, LGG, BRCA, THCA, PAAD, STAD, COAD, OV), les données d'ARNm permettent d'obtenir un C-index médian significativement plus important que celui obtenu avec les données miRNA-seq (p-valeur < 0,05, test de Wilcoxon avec correction de Benjamini-Hochberg).

Enfin, pour ESCA, les données de miARN ont une plus grande valeur prédictive de la survie que les données d'ARNm (*i.e.* C-index médian significativement plus important, p-valeur < 0,01, test de Wilcoxon avec correction de Benjamini-Hochberg). L'intérêt des miARN pour prédire la survie dans ce cancer a déjà été démontré [MATHÉ et collab., 2009; YANG et collab., 2020]. De plus, il est intéressant de remarquer que le C-index médian obtenu avec les données mRNA-seq (0.51) n'est pas significativement différent de 0,5 (p-valeur > 0,05, test de Wilcoxon unilatéral), alors celui obtenu avec les données miRNA-seq (0,61) l'est (p-valeur < 0.001, test de Wilcoxon unilatéral). Ce résultat suggère que les miARN pourraient avoir un rôle biologique particulier pour ce cancer.

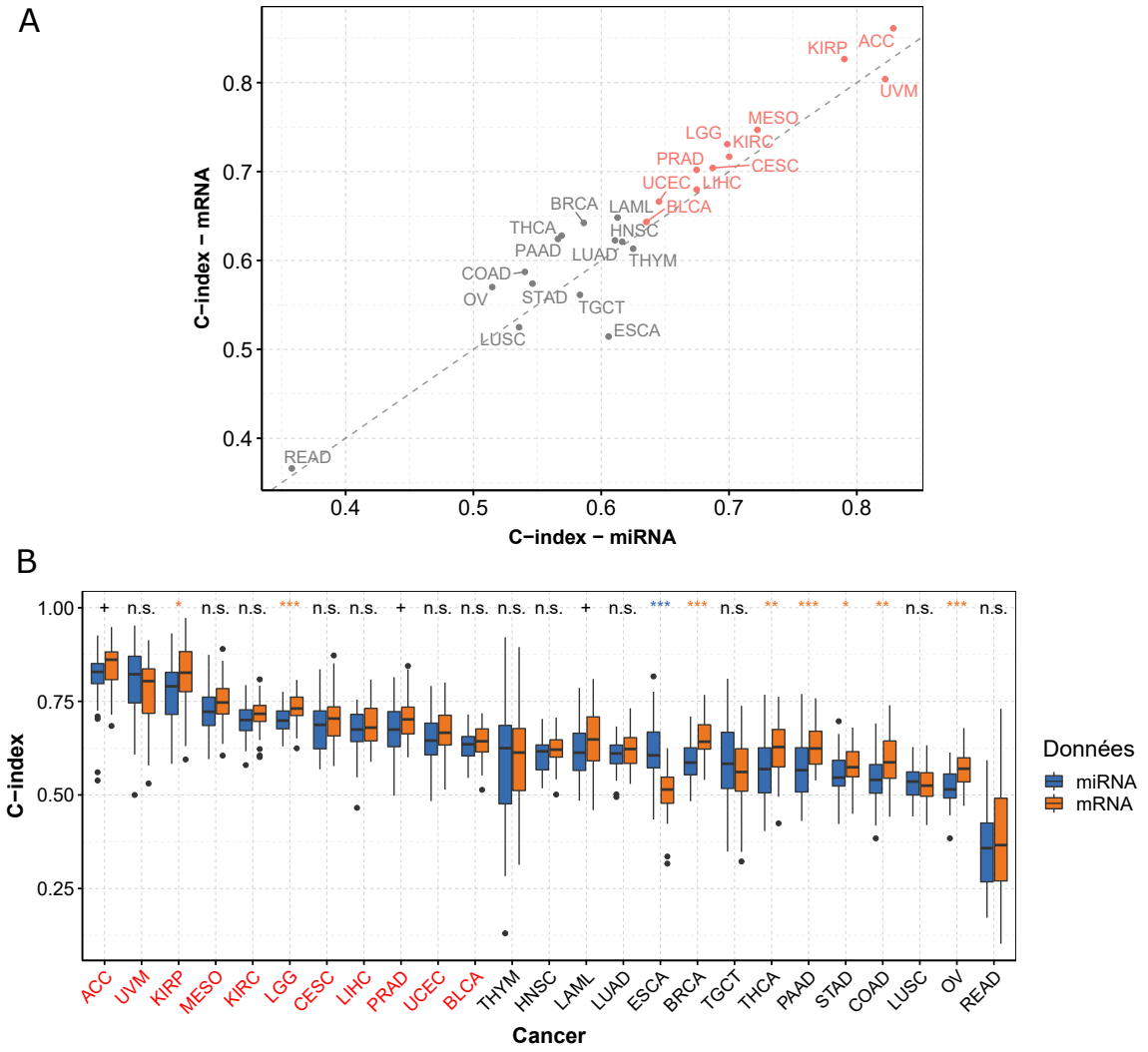


FIGURE 4.4 – **C-index obtenus avec les données RNA-seq pour les ARNm et les miARN pour 25 cancers de TCGA**

Les 11 cancers dont le nom est écrit en rouge sont ceux qui seront étudiés par la suite.
(A) C-index médians obtenus avec les données mRNA-seq en fonction de ceux obtenus avec les données miRNA-seq pour 25 cancers de TCGA. La droite d'équation $y = x$ est tracée en pointillé.
(B) *Boxplot* des C-index obtenus avec les données RNA-seq pour les miARN (bleu) et les ARNm (orange) pour 25 cancers de TCGA. Les p-valeurs d'un test de Wilcoxon corrigées par la méthode de Benjamini-Hochberg sont indiquées au sommet du graphique sous forme d'étoiles (légende ci-dessous). La p-valeur est bleue lorsque le C-index médian est significativement plus important pour les miARN, orange lorsque le C-index médian est significativement plus important pour les ARNm, et noir lorsque la différence n'est pas significative (le seuil de 0,05 a été choisi).
n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

4.4 Prédiction de la survie avec les variables cliniques et les données miRNA-seq

Pour étudier l'intérêt des miARN pour la prédiction de la survie, nous utilisons la même démarche décrite à la partie 2.3 pour les données mRNA-seq. Pour rappel, nous calculons 50 C-index obtenus avec six variables cliniques classiques, 50 C-index obtenus avec les données miRNA-seq, et 50 C-index en combinant les données cliniques et les données miRNA-seq. Nous calculons ces 50 C-index par 10 répétitions d'une validation croisée (K=5) dans chacun des cas. Les six variables cliniques classiques que nous utilisons, lorsque celles-ci sont disponibles, sont l'âge, le genre, le grade, et les nomenclatures T, N, et M [Lu et collab., 2018; Milanez-Almeida et collab., 2020; Rosenberg et collab., 2005]. Nous avons décrit ces variables cliniques dans la partie 2.3.

Nous retenons une donnée clinique uniquement lorsqu'elle est présente pour au moins 90% des patients de TCGA. L'âge est disponible pour les 11 cancers, le genre pour les 8 cancers non unisexe (*i.e.* CESC et UCEC sont des cancers touchant uniquement les femmes, et PRAD uniquement les hommes), le grade n'est pas disponible pour ACC, UVM, KIRP et MESO, et les variables T, N, et M ne sont pas disponibles pour LGG, CESC, UCEC. Pour ACC, T et N sont disponibles, mais pas M. Pour PRAD, T est disponible, mais pas N et M (Tableau 4.2).

Tableau 4.2 – **Variables cliniques présentes pour chacun des 11 cancers étudiés.**
1 : la variable clinique est présente pour au moins 90% des patients ; 0 : la variable clinique n'est pas présente pour au moins 90% des patients.

Cancer	Age	Genre	Grade	T	N	M
ACC	1	1	0	1	1	0
UVM	1	1	0	1	1	1
KIRP	1	1	0	1	1	1
MESO	1	1	0	1	1	1
KIRC	1	1	1	1	1	1
LGG	1	1	1	0	0	0
CESC	1	0	1	0	0	0
LIHC	1	1	1	1	1	1
PRAD	1	0	0	1	0	0
UCEC	1	0	1	0	0	0
BLCA	1	1	1	1	1	1

Ensuite, pour combiner les données cliniques et les données miRNA-seq, nous avons utilisé les indices pronostiques calculés avec les données miRNA-seq comme septième co-variable explicative, en plus des variables cliniques. Pour évaluer la valeur ajoutée des données miRNA-seq dans la prédiction par rapport aux données cliniques seules, nous avons effectué un test de Wilcoxon unilatéral pour chaque cancer. Ce test permet d'observer si le C-index médian obtenu avec la combinaison des données cliniques et miRNA-seq est significativement supérieur à celui obtenu avec uniquement les données cliniques.

Nous pouvons ainsi identifier les cancers pour lesquels le profil d'expression des miARN d'une biopsie permet d'améliorer les prédictions obtenues avec quelques variables cliniques classiques. Nous corrigeons les p-valeurs suivant la méthode de Benjamini-Hochberg pour l'ensemble des 11 cancers.

Pour 7 cancers (ACC, UVM, MESO, LGG, CESC, LIHC, PRAD) sur les 11 étudiés, l'ajout des données miRNA-seq aux données cliniques classiques améliorent le C-index significativement par rapport aux données cliniques seules (p-valeur < 0,05, test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg) (Fig. 4.5). En revanche, pour KIRP, KIRC, UCEC, et BLCA, les données miRNA-seq n'apportent pas de valeur prédictive par rapport aux données cliniques. Cependant, comme nous l'avons fait remarquer à la partie 2.3, au-delà de la simple prédiction, les données miRNA-seq peuvent être utilisées pour différentes tâches :

— stratifier les patients suivant les profils transcriptomiques [RICKETTS et collab., 2018].

— identifier des marqueurs prédictifs de réponse aux traitements [TERNÈS et collab., 2017].

— identifier de potentielles cibles thérapeutiques [WEI et collab., 2017].

Nous étudierons donc l'ensemble des 11 cancers par la suite.

Enfin, il est intéressant de remarquer que les C-index obtenus sont élevés pour certains cancers et faibles pour d'autres (*e.g.* C-index médian de 0,81 pour KIRP, et de 0,51 pour MESO, Fig. 4.5), et que cette qualité de prédiction ne dépend pas de façon claire du nombre de paramètres cliniques disponibles. En effet, pour CESC et UCEC, seuls le grade et l'âge sont disponibles, et les C-index médians obtenus sont de 0,56 et 0,69 respectivement. De plus, malgré la présence des données T, N et M pour MESO, le C-index médian obtenu avec les données cliniques est seulement de 0,51. La classification TNM qui définit le stade du cancer ne permet donc pas prédire la survie pour ce cancer. Ce résultat a déjà été observé dans d'autres études [MILANEZ-ALMEIDA et collab., 2020], et sur d'autres jeux de données [ZHUO et collab., 2019].

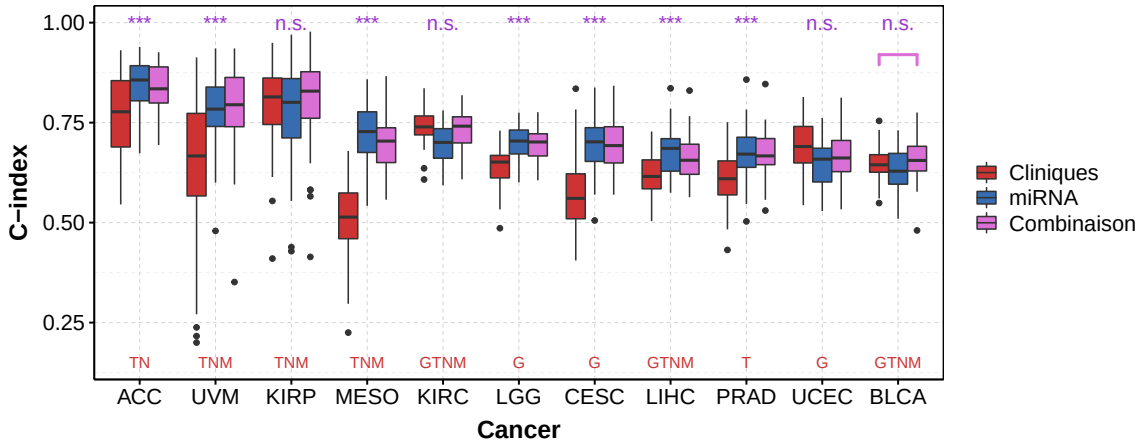


FIGURE 4.5 – **C-index obtenus avec les données cliniques, miRNA-seq, et en combinant les deux types de données pour les 11 cancers étudiés.**
Pour chaque cancer, 50 C-index sont calculés en utilisant 6 variables cliniques (âge, genre, grade, T, N, M - rouge) lorsqu'elles sont disponibles, les données miRNA-seq (bleu), et en combinant les deux types de données (*i.e.* les 6 variables cliniques et l'indice pronostique calculé avec les données miRNA-seq sont utilisés - violet). Pour chaque cas, les 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5).
L'âge est disponible pour les 11 cancers, le genre pour les 8 cancers non unisexe (*i.e.* CESC et UCEC sont des cancers touchant uniquement les femmes, et PRAD uniquement les hommes). Les lettres rouges en bas du graphique indique la présence ou non des autres variables cliniques (G : grade, T : « *Tumor* », N : « *Node* », M : « *Metastasis* »).
Les p-valeurs corrigées par la méthode de Benjamini-Hochberg d'un test de Wilcoxon unilatéral permettant d'observer si le C-index médian obtenu avec la combinaison des données cliniques et miRNA-seq (violet) est significativement supérieur à celui obtenu avec uniquement les données cliniques (rouge) sont indiquées en violet sous forme d'étoiles suivant le niveau de significativité (légende ci-dessous).
n.s. : non-significatif; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

4.5 Dégradation des données de séquençage

4.5.1 Dégradation de la profondeur de séquençage

Afin d'évaluer l'effet de la profondeur de séquençage sur la qualité de prédiction de la survie des patients avec un modèle de Cox, nous avons mesuré les métriques d'évaluation des prédictions (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS) après dégradation de la profondeur de séquençage. Pour dégrader les données de séquençage, nous avons utilisé une méthode de sous-échantillonnage [ROBINSON et STOREY, 2014]. Le paramètre clef de cette méthode est la proportion de sous-échantillonnage, $\varepsilon \in]0,1]$. Ce paramètre permet de calibrer l'intensité de sous-échantillonnage, et donc de dégradation des données. Pour chaque donnée de comptage (nombre de lectures) R_{ij} obtenue pour un patient i et un gène j, une donnée de comptage sous-échantillonnée d'une proportion ε, notée $\tilde{R}_{ij}$, est tirée suivant une loi binomiale de paramètres R_{ij} (nombre d'épreuves) et ε

(probabilité de succès) :

$$\tilde{R}_{ij} \sim \text{Binom}(R_{ij}, \varepsilon), \tag{4.1}$$

pour chaque patient $i = 1, \ldots, n$ et chaque gène $j = 1, \ldots, p$.

Ainsi, plus le paramètre ε est proche de 0, plus les données de comptage RNA-seq seront dégradées : une proportion de ε (*e.g.* 0,01) correspond à une dégradation des données de séquençage d'un facteur $\delta = \frac{1}{\varepsilon}$ (*e.g.* 100). Dans cette étude, nous choisissons 1 (pas de dégradation), 10, 100, 1 000 et 10 000 comme facteurs de dégradation δ possibles. Le calcul des métriques d'évaluation de la qualité de prédiction (*i.e.* C-index, IBS, p-valeur du modèle de Cox univarié) pour différents facteurs de dégradation permet de calculer des courbes de saturation [Bass et collab., 2019; Tarazona et collab., 2011]. Ces courbes se définissent comme la métrique d'évaluation des performances du modèle en fonction du facteur de dégradation δ.

4.5.2 Diminution du nombre de patients

Pour étudier l'impact du nombre de patients sur les capacités de prédiction, nous avons diminué le pourcentage x de patients du jeu de données d'apprentissage. Dans cette étude, nous avons choisi $x = 10, 20, \ldots, 80\%$. En revanche, pour que les métriques d'évaluation du modèle (*i.e.* C-index, IBS, p-valeur du modèle de Cox univarié) ne soient pas biaisées, le jeu de données de test est toujours composé de 20% des patients.

Nous avons décrit la procédure permettant d'évaluer l'impact de la dégradation de la taille des banques et du nombre de patients du jeu de données d'apprentissage sur la Figure 4.6. Le principe est le même que celui décrit en introduction (partie 1.7.2), mais un sous-échantillonnage des données de comptage et du pourcentage de patients définissant le jeu de données d'apprentissage est ajouté (en caractères gras sur la Figure 4.6).

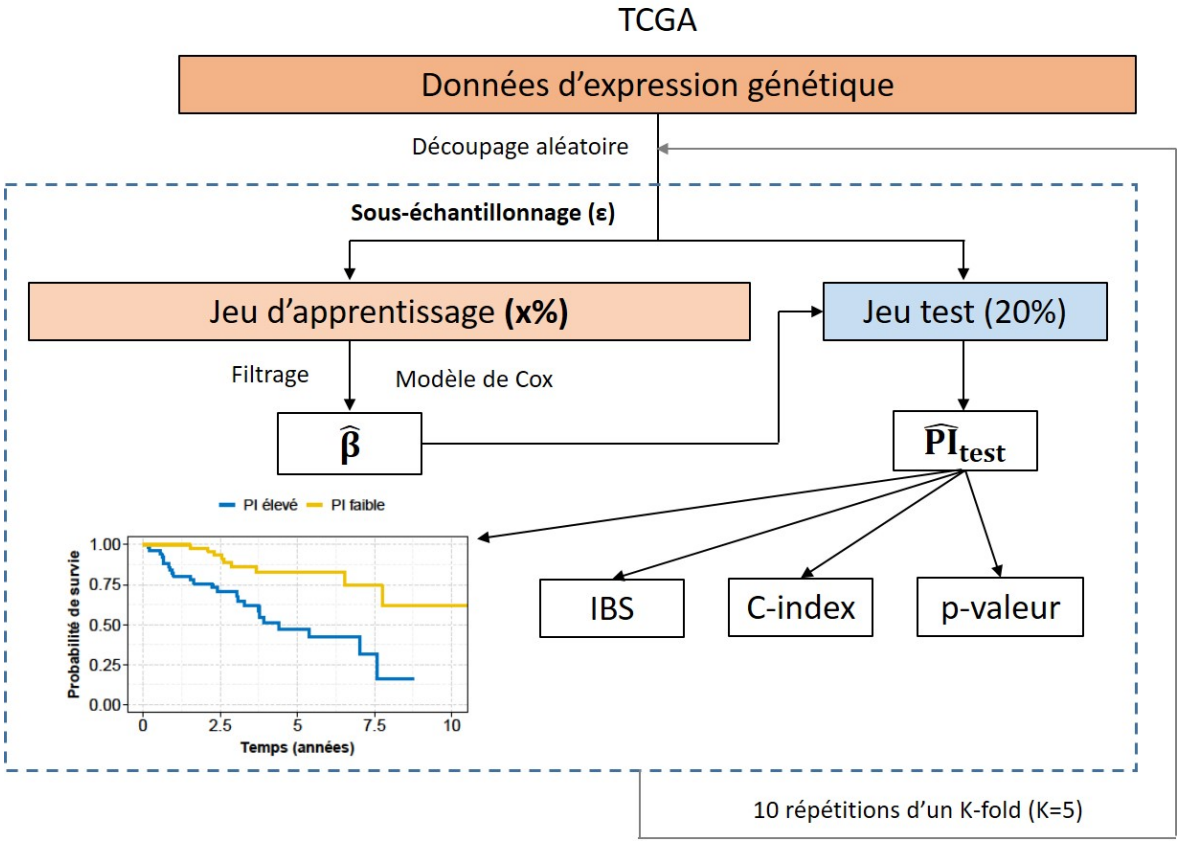


FIGURE 4.6 – **Procédure d'évaluation de l'impact de la dégradation des données de comptage RNA-seq et du nombre de patients sur la prédiction.**
Identique à la Figure 1.6 avec les "dégradations" appliquées en gras. Les données de comptage RNA-seq sont sous-échantillonnées ($\varepsilon = 0,0001; 0,001; 0,01; 0,1; 1$), le modèle de Cox est entraîné sur x% des patients ($x = 10, 20, \ldots, 80$), et l'estimateur $\hat{\boldsymbol{\beta}}$ du modèle de Cox pénalisé est calculé. Ce vecteur permet d'estimer le vecteur des indices pronostiques $\widehat{\mathbf{PI}}_{\mathbf{test}}$ du jeu de données test, et les différents indicateurs d'évaluation de la qualité de prédiction (*i.e.* score de Brier intégré - IBS, C-index, p-valeur du modèle de Cox univarié). Ce processus est répété 50 fois (10 répétitions d'une validation croisée, K=5).

4.6 Impact de la taille des banques et du nombre de patients sur la prédiction

Tout d'abord, pour l'ensemble des onze cancers étudiés et pour les trois métriques d'évaluation des prédictions (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS), nous observons que les capacités de prédiction augmentent avec le nombre de patients du jeu d'apprentissage et la profondeur de séquençage (Fig. 4.7 pour KIRC, et données non montrées pour les autres cancers). Dans les deux parties suivantes, nous analyserons en détails l'impact de la profondeur de séquençage et du nombre de patients sur les prédictions.

4.6.1 Impact de la taille des banques

Pour KIRC, la baisse du C-index devient significative dès une dégradation des données de séquençage d'un facteur 10 (p-valeur < 0,05, test de Wilcoxon unilatéral, Fig. 4.7). Des résultats similaires sont obtenus pour CESC et LIHC (p-valeur < 0,05, test de Wilcoxon unilatéral), et dans une moindre mesure pour MESO et PRAD (p-valeur < 0,1, test de Wilcoxon unilatéral) (données non montrées). Il n'est donc pas possible de diminuer la taille des banques des miARN d'un facteur 10 sans détériorer significativement le C-index obtenu avec un modèle de Cox et une pénalisation ridge pour ces cinq cancers. Notons que cette conclusion diffère de celle observée avec les ARNm pour lesquels une dégradation d'un facteur 100 n'affecte pas le C-index de manière significative pour ces cinq cancers. En revanche, pour les six autres cancers (*i.e.* ACC, UVM, KIRP, LGG, UCEC, BLCA), la profondeur de séquençage peut être diminué d'un facteur 10 au moins sans affecter de manière significative le C-index (p-valeur ≥ 0,05, test de Wilcoxon unilatéral).

Quelques différences sont à noter lorsque l'on choisit une autre métrique d'évaluation des prédictions. Par exemple, pour PRAD, il est possible de dégrader les données de séquençage d'un facteur 10 sans dégrader significativement la p-valeur du modèle de Cox univarié, alors que le C-index est significativement diminué (p-valeur < 0,1, test de Wilcoxon unilatéral). De plus, en considérant l'IBS comme métrique d'évaluation, la taille des banques peut être diminuée d'un facteur 10 au moins pour l'ensemble des cancers, sauf pour BLCA (p-valeur < 0,1, test de Wilcoxon unilatéral).

Ainsi, dans la suite, nous considérerons que la taille des banques peut être diminuée d'un facteur de dégradation δ si et seulement si aucune des trois métriques d'évaluation (*i.e.* C-index, IBS, p-valeur du modèle de Cox univarié) n'est dégradée au niveau de significativité de 0,1 pour ce facteur δ. Ainsi, en considérant ce critère, la taille des banques ne peut pas être réduite d'un facteur 10 pour ACC, MESO, KIRC, CESC, LIHC, PRAD et BLCA. Pour affiner l'étude de ces sept cancers, nous avons dégradé la taille des banques d'un facteur 5 et 2 afin d'augmenter la granularité. Comme décrit précédemment, nous avons ensuite effectué des tests de Wilcoxon unilatéraux pour tester si la baisse d'au moins une métrique d'évaluation est significative au niveau 0,1 pour ces deux facteurs de dégradation.

Le tableau 4.3 résume le facteur de dégradation maximum des données miRNA-seq et la taille médiane de la banque conseillée pour minimiser la profondeur de séquençage - donc les coûts - tout en obtenant une qualité de prédiction optimale. Pour ACC, PRAD et BLCA, la taille médiane des banques requises est de l'ordre de 1 million de lectures ; pour LGG, elle est de l'ordre de 745 000 lectures ; pour KIRP, KIRC et UCEC, elle se situe autour de 500 000 lectures. De manière intéressante, pour UVM, une dégradation des données de séquençage d'un facteur 1 000 ne dégrade pas les prédictions obtenues, et une taille médiane des banques de seulement 5 000 lectures permet déjà d'obtenir des résultats optimaux. En revanche, pour MESO, CESC et LIHC, la taille des banques ne peut pas être

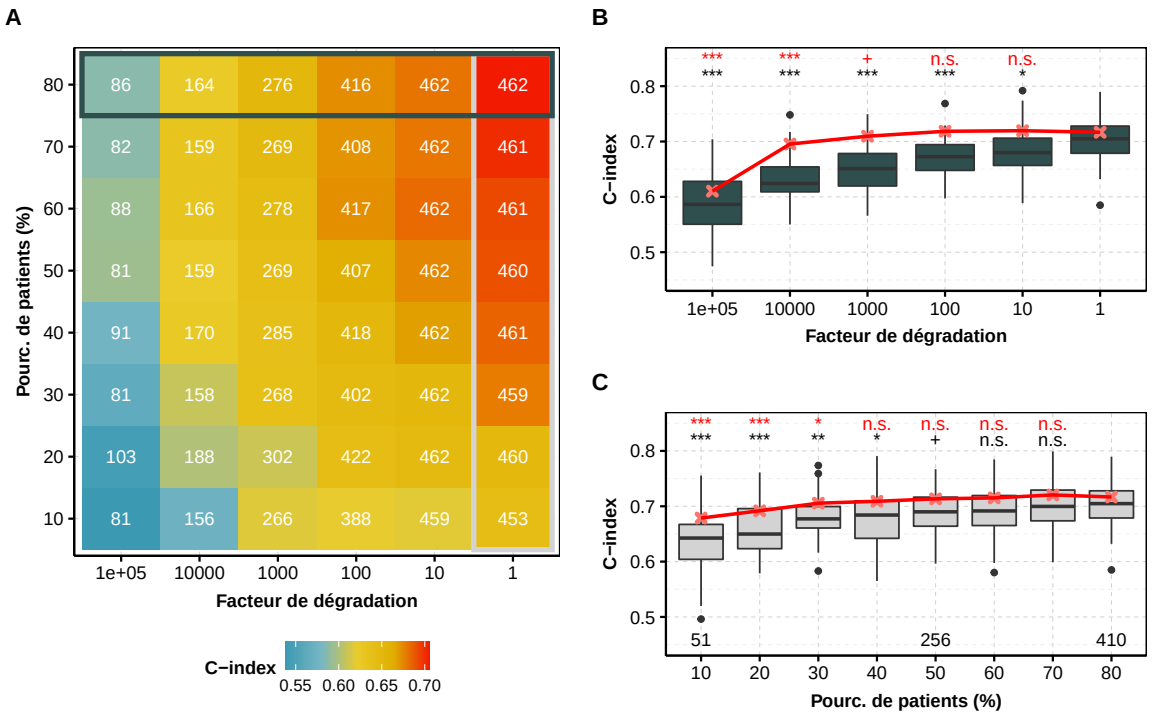


FIGURE 4.7 – **Impact de taille des banques et du nombre de patients du jeu d'apprentissage sur le C-index pour KIRC**

(A) C-index médians calculés pour différents facteurs de dégradation δ de la taille des banques ($\delta = 1$: pas de dégradation, $\delta > 1$: dégradation de la donnée de comptage R_{ij} suivant une loi binomiale de paramètre R_{ij} et $\varepsilon = 1/\delta$ donnée en (4.1)) et différents pourcentage de patients définissant le jeux de données d'apprentissage. Les facteurs de dégradation sont 1, 10, ..., 100 000 (abscisse), et les pourcentages de patients définissant le jeu de données d'apprentissage sont 10, 20, ..., 80 (ordonnée). L'axe des abscisses est volontairement inversé pour être croissant en profondeur de séquençage (*i.e.* plus δ est faible, plus la profondeur de séquençage est importante).

(B) *Boxplot* des C-index calculés pour chaque facteur de dégradation. Pour chaque cas, le jeu de données d'apprentissage contient 80% des patients (ligne supérieure encadrée de gris foncé sur la Figure A). Là encore, l'axe des abscisses est inversé, comme pour la figure A. Les C-index médians obtenus avec données mRNA-seq sont indiqués par des croix rouges reliées par des segments.

(C) *Boxplot* des C-index calculés pour chaque pourcentage de patients définissant le jeu de données d'apprentissage. Pour chaque cas, les données RNA-seq ne sont pas dégradées (dernière colonne encadrée de gris claire sur la Figure A). Les C-index médians obtenus avec données mRNA-seq sont indiqués par des croix rouges reliées par des segments.

Pour chaque scénario, 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5), et un test de Wilcoxon unilatéral permet de comparer le C-index médian calculé sans dégradation (profondeur de séquençage non dégradée pour la Figure B, et 80% des patients utilisés dans le jeu d'apprentissage pour la Figure C) et après dégradation.

n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

diminué, même d'un facteur 2, sous peine de détériorer les prédictions du modèle de Cox avec pénalisation ridge. Ainsi, pour ces trois cancers, la taille moyenne des banques doit au moins être de 5 millions de lectures, et les prédictions pourraient potentiellement être améliorées en augmentant la profondeur de séquençage.

TABLEAU 4.3 – **Facteur de dégradation maximum et taille de banque médiane (en milliers) conseillée pour chacun des 11 cancers pour les données RNA-seq de miARN.**
Le signe « < » pour le facteur de dégradation et « > » pour la taille médiane de la banque signifie que les données RNA-seq ne peuvent pas être sous-échantillonnées. Les trois cancers concernés (MESO, CESC et LIHC) sont notés en gras.

Cancer	ACC	UVM	KIRP	**MESO**	KIRC	LGG	**CESC**	**LIHC**	PRAD	UCEC	BLCA
Facteur de dégradation δ	5	1000	10	<1	5	10	<1	<1	5	10	5
Taille médiane de banque conseillée (milliers de lectures)	1000	5	610	>5200	500	750	>4800	>4800	860	430	970

En revanche, pour les ARNm, nous partageons les conclusions de MILANEZ-ALMEIDA et collab. [2020] : la taille des banques des données mRNA-seq peut être diminuée d'un facteur 100 sans affecter les capacités de prédiction de manière significative. Cette diminution correspond en moyenne à des tailles de banques de l'ordre de 500 000 lectures (Tableau 4.4). Deux exceptions sont à noter :

— pour UCEC, la profondeur de séquençage ne peut être diminué que d'un facteur 10, ce qui correspond à une taille de banque médiane de l'ordre de 2,5 millions de lectures.

— pour BLCA, la profondeur de séquençage peut être diminué d'un facteur 1 000, ce qui correspond à une taille de banque médiane de l'ordre de 50 000 lectures.

TABLEAU 4.4 – **Facteur de dégradation maximum et taille de banque médiane (en milliers) conseillée pour chacun des 11 cancers pour les données mRNA-seq.**

Cancer	ACC	UVM	KIRP	MESO	KIRC	LGG	CESC	LIHC	PRAD	UCEC	BLCA
Facteur de dégradation	100	100	100	100	100	100	100	100	100	10	1000
Taille médiane de banque conseillée (milliers de lectures)	380	420	420	470	500	470	480	460	450	2300	41

Une réduction de la taille des banques permet de diminuer les coûts. Ainsi, à budget constant, une telle diminution permet de séquencer un nombre plus important d'échantillons. De multiples perspectives peuvent alors être envisagées :

1. l'étude de l'hétérogénéité intra-tumorale (séquençage de différentes parties de la tumeur).

2. un séquençage d'échantillons de tumeur obtenus par biopsies à différents instants t afin d'étudier l'effet d'un traitement sur l'évolution du transcriptome de la tumeur en fonction du temps.
3. une augmentation de la taille de la cohorte de patients.

Ces différentes perspectives peuvent être associées afin de répondre à la problématique considérée. Les deux premières perspectives ont des intérêts biologiques et cliniques évidents [GERLINGER et collab., 2012; MCGRANAHAN et SWANTON, 2017], et l'utilité de l'augmentation de la taille de la cohorte de patients sera détaillée dans le paragraphe suivant.

4.6.2 Impact du nombre de patients

Pour l'ensemble des cancers, les prédictions obtenues avec le modèle de Cox et une pénalisation ridge sont saturées en fonction du nombre de patients à la fois pour les miARN et les ARNm (Fig. 4.7.C). Ainsi, si l'on se limite à ce modèle, augmenter la taille de la cohorte ne permettra pas d'obtenir de meilleures prédictions. En revanche, un nombre plus important de patients permettrait :

1. d'utiliser des modèles mathématiques plus élaborés et nécessitant un nombre important de patients (*e.g.* réseaux de neurones profonds).
2. de stratifier les patients en fonction de caractéristiques cliniques (*e.g.* sexe, âge), phénotypiques (*e.g.* grade, stade, pureté de la tumeur), ou des données transcriptomiques.
3. d'augmenter la dimension de la matrice de co-variables en intégrant des termes d'interactions dans l'analyse (*e.g.* interactions entre les gènes, interactions entre gènes et variables cliniques, etc.).

4.7 Cause de la dégradation de la qualité de prédiction

4.7.1 Deux hypothèses envisagées

Tout d'abord, rappelons que nous définissons un gène comme « détecté » si son niveau d'expression CPM-normalisé est supérieur à 1 pour au moins 1% des patients du jeu de données d'apprentissage (partie 1.4.1). Deux hypothèses peuvent être émises pour expliquer la diminution des capacités prédictives induite par la dégradation de la qualité du séquençage :

1. lorsque les données de comptage miRNA-seq sont dégradées, le nombre de gènes détectés décroît (Fig. 4.8.A) [MORTAZAVI et collab., 2008; SIMS et collab., 2014], et seul le niveau d'expression des gènes les plus fortement exprimés peut-être mesuré (Fig. 4.8.B). Les mêmes conclusions sont observés pour les données miRNA-seq

pour les autres cancers (données non montrées), et pour les données mRNA-seq (Fig. Annexe A.14). Cependant, les gènes qui ont un niveau d'expression trop faible pour être mesuré correctement lorsque la taille des banques diminue peuvent être des co-variables à forte valeur prédictive. Ainsi, les prédictions peuvent être dégradées par l'absence de ces co-variables parmi les prédicteurs.

2. lorsque les données de comptage miRNA-seq sont dégradées, le rapport signal sur bruit diminue, et la qualité des données est moins bonne. Pour illustrer cette hypothèse, supposons que la donnée de comptage RNA-seq R est déterministe, et que le bruit de mesure η suit une loi normale $N(0, \sigma^2)$. Le signal initial est alors $S = R + \eta$, et le rapport signal sur bruit est définie par $\frac{E(S^2)}{E(\eta^2)}$. Après dégradation de R d'un facteur ε, le nouveau signal est $\tilde{S} = \varepsilon R + \eta$. Ainsi, le rapport signal sur bruit de S est $RSB(S) = \frac{R^2+\sigma^2}{\sigma^2}$, et celui de $\tilde{S}$ est $RSB(\tilde{S}) = \frac{\epsilon^2 R^2+\sigma^2}{\sigma^2}$. Le rapport signal sur bruit a donc bien diminuer après dégradation.

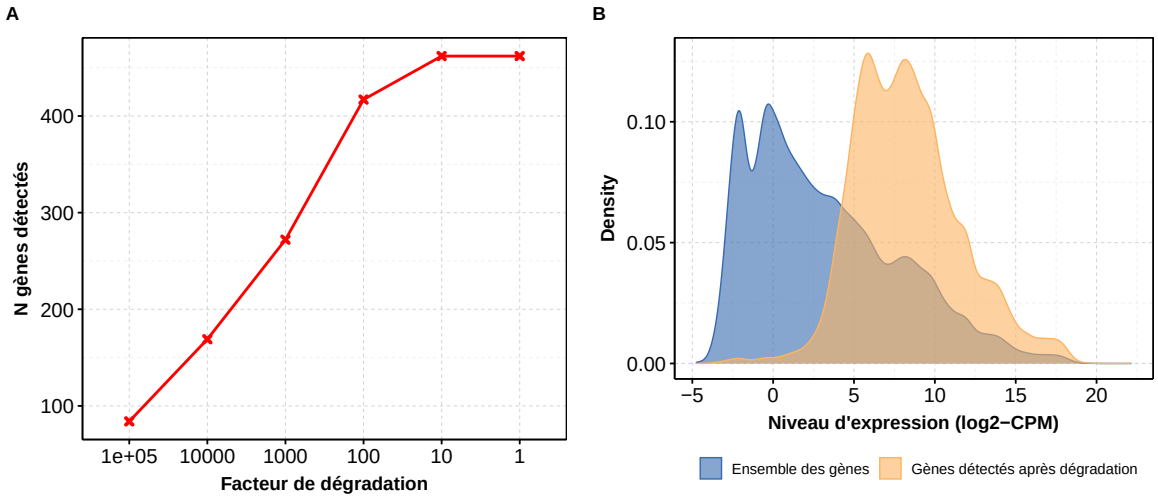


FIGURE 4.8 – **Impact de la dégradation de la taille des banques des données miRNA-seq sur le nombre de gènes détectés (A) et leur niveau d'expression (B) pour KIRC .**
(A) Les données brutes sont dégradées suivant 5 facteurs différents (1, 10, 100, 1 000, 10 000, 100 000 - abscisse), et le nombre de gènes détectés est calculé (ordonnée). (B) Densité (estimation de la distribution) du niveau d'expression de l'ensemble des gènes (bleu), et des gènes détectes après dégradation d'un facteur 10 000 (jaune) pour l'ensemble des patients. Les données de comptage sont normalisées (log2-CPM) dans les deux cas.

4.7.2 Test de la première hypothèse

Pour tester la première hypothèse, nous avons comparé les C-index obtenus avec l'ensemble des miARN, et ceux obtenus uniquement avec les gènes les plus exprimés. Cela correspond en moyenne à une diminution d'un facteur 2 du nombre de prédicteurs dans le modèle de Cox (*i.e.* 210 gènes en moyenne détectés après dégradation des données

miRNA-seq d'un facteur 10 000 pour l'ensemble des cancers étudiés). Dans ces deux scénarios, les données de séquençage ne sont pas dégradées, mais seul le nombre de gènes pris en compte diffère. Les C-index obtenus avec tous les gènes sont représentés par des boites bleues sur la Figure 4.9, et ceux obtenus avec uniquement les gènes les plus exprimés sont représentés par des boites jaunes. Nous avons utilisé un test de Wilcoxon unilatéral pour tester si le C-index médian obtenu en utilisant tous les miARN est supérieur à celui obtenu uniquement avec les 210 miARN les plus exprimés sans dégradation. Les p-valeurs corrigées par la méthode de Benjamini-Hochberg (BENJAMINI et HOCHBERG [1995], partie 1.5.3) sont indiquées en bleu sous forme d'étoiles sur la Figure 4.9 suivant le niveau de significativité.

Pour les miARN, réduire la dimension en ne gardant que les 210 gènes les plus exprimés n'impacte pas significativement les capacités de prédiction, excepté pour CESC (p-valeur < 0,001, test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg), et dans une moindre mesure pour MESO (p-valeur < 0,1, test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg) (Fig. 4.9). Pour ces deux cancers, les 210 gènes les plus exprimés ne suffisent pas à prédire aussi bien la survie que si l'ensemble des miARN étaient utilisés. En revanche, pour les 9 autres cancers, la première hypothèse émise dans la partie précédente n'est pas vérifiée : la diminution des capacités prédictives du modèle n'est pas due au nombre moins important de gènes détectés et utilisés comme prédicteur dans le modèle de Cox. De plus, cela montre que les 210 miARN les plus exprimés suffisent à prédire la survie des patients. En d'autres termes, les 300 miARN les moins exprimés n'apportent pas de valeur ajoutée pour prédire la survie au 210 miARN les plus exprimés dans le modèle de Cox.

Nous avons utilisé la même démarche pour les ARNm (Fig. Annexe A.15). En moyenne les 11 200 gènes les plus exprimés détectés après une dégradation des données d'un facteur 10 000 sont utilisés (boites jaunes), et ils suffisent à obtenir des capacités de prédiction comparable au scénario où l'ensemble des gènes est utilisé (18 000 en moyenne, boites bleues). Ainsi, la première hypothèse émise dans la partie précédente n'est vérifiée pour aucun des onze cancers pour les données mRNA-seq.

4.7.3 Test de la deuxième hypothèse

Pour tester la deuxième hypothèse, nous avons calculé les C-index obtenus après dégradation d'un facteur 10 000 (*i.e.* 210 gènes en moyenne sont détectés, et les données de comptage sont dégradés), et ceux obtenus avec les mêmes 210 gènes (en moyenne) mais sans dégradation. Ces deux scénarios correspondent aux boites vertes et jaunes de la Figure 4.9, respectivement. Un test de Wilcoxon unilatéral est de nouveau utilisé pour tester si le C-index médian obtenu en utilisant uniquement les 210 miARN les plus exprimés sans dégradation est supérieur à celui obtenu après dégradation d'un facteur 10 000. Les p-valeurs corrigées par la méthode de Benjamini-Hochberg (BENJAMINI et HOCHBERG

[1995], partie 1.5.3) sont indiquées en vert sous forme d'étoiles sur la Figure 4.9 suivant le niveau de significativité.

Pour l'ensemble des onze cancers, le C-index médian obtenu après dégradation est significativement inférieur à celui obtenu sans dégradation mais avec les même prédicteurs (p-valeur < 0,05 pour UVM et p-valeur < 0,001 pour les dix autres cancers, test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg). La deuxième hypothèse émise précédemment est donc vérifiée : le sous-échantillonnage induit une diminution du rapport signal sur bruit des données de comptage RNA-seq et donc une dégradation de la qualité des données et de la prédiction.

Nous avons utilisé la même démarche pour les ARNm (Fig. Annexe A.15), et les mêmes conclusions peuvent être tirées de cette analyse. Pour l'ensemble des onze cancers, le C-index médian obtenu après dégradation est significativement inférieur à celui obtenu sans dégradation mais avec les même prédicteurs (p-value < 0,05 pour les onze cancers, test de Wilcoxon unilatéral avec correction de Benjamini-Hochberg).

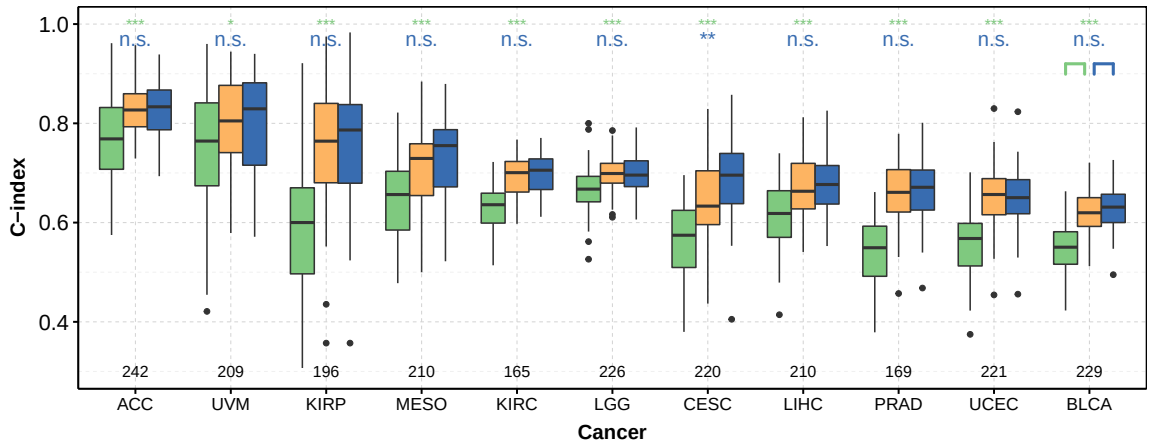


FIGURE 4.9 – **Evaluation des deux hypothèses pour expliquer la diminution des capacités de prédiction induite par la dégradation des données miRNA-seq des 11 cancers étudiés.**

Boxplot des C-index obtenus avec les données miRNA-seq pour un facteur de dégradation de 10 000 (vert), sans dégradation (bleu), et sans dégradation mais avec uniquement les gènes détectés après dégradation d'un facteur 10 000 (jaune). Le nombre de gènes détectés après dégradation d'un facteur 10 000 est indiqué en noir en bas du graphique pour chaque cancer. Pour chaque scénario, 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5).

Les p-valeurs corrigées par la méthode de Benjamini-Hochberg d'un test de Wilcoxon unilatéral permettant de tester si le C-index médian obtenu dans le « scénario jaune » est significativement plus grand que celui obtenu dans le « scénario vert » sont indiquées sous forme d'étoiles vertes en haut du graphique.

Les p-valeurs corrigées par la méthode de Benjamini-Hochberg d'un test de Wilcoxon unilatéral permettant de tester si le C-index médian obtenu dans le « scénario bleu » est significativement plus grand que celui obtenu dans le « scénario jaune » sont indiquées sous forme d'étoiles bleues en haut du graphique.

n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

4.8 Conclusions

Tout d'abord, nous avons observé que la taille des banques varie suivant les patients. Cette variation est due aux biais techniques induits par les différentes étapes du séquençage.

Ensuite, nous avons montré que le C-index médian obtenu avec les ARNm est équivalent à celui obtenu avec les miARN pour 16 cancers (ACC, UVM, MESO, KIRC, CESC, LIHC, PRAD, UCEC, BLCA, THYM, HNSC, LAML, LUAD, TGCT, LUSC, READ), significativement supérieur pour 8 d'entre eux (KIRP, LGG, BRCA, THCA, PAAD, STAD, COAD, OV), et significativement inférieur pour ESCA. De plus, nous avons observé une forte corrélation entre les C-index médians obtenus avec les données RNA-seq d'ARNm et de miARN. Les prédictions dépendent donc plus du cancer étudié que du type de données utilisées.

Troisièmement, nous avons montré que pour 7 des 11 cancers étudiés, l'ajout des données miRNA-seq aux données cliniques classiques permet d'augmenter les capacités de prédiction. Pour les 4 autres cancers, d'autres applications sont envisageables (*e.g.* détection de biomarqueurs prédictifs de la réponse aux traitements), et ils méritent aussi d'être étudiés.

Ensuite, nous partageons les conclusions de Milanez-Almeida et collab. [2020] : la taille des banques des données mRNA-seq peut être diminuée d'un facteur 100 sans affecter les capacités de prédiction de manière significative. Cette diminution correspond en moyenne à des tailles de banques de l'ordre de 500 000 lectures. Deux exceptions sont cependant à noter : pour UCEC, la profondeur de séquençage ne peut être diminué que d'un facteur 10, ce qui correspond à une taille de banque médiane de l'ordre de 2,5 millions de lectures ; et pour BLCA, la profondeur de séquençage peut être diminué d'un facteur 1000, ce qui correspond à une taille de banque médiane de l'ordre de 50 000 lectures.

Cinquièmement, nous avons vu que pour les miARN, les résultats sont plus hétérogènes suivant les cancers. Pour ACC, PRAD et BLCA, la taille médiane des banques requises est de l'ordre de 1 million de lectures ; pour LGG, elle est de l'ordre de 745 000 lectures ; pour KIRP, KIRC et UCEC, elle se situe autour de 500 000 lectures. De manière intéressante, pour UVM, une dégradation des données de séquençage d'un facteur 1 000 ne dégrade pas les prédictions obtenues, et une taille médiane des banques d'uniquement 5 000 lectures permet de converger vers des prédictions optimales. En revanche, pour MESO, CESC et LIHC, la taille des banques ne peut pas être diminuée d'un facteur 2, sous peine de détériorer les prédictions du modèle de Cox avec pénalisation ridge. Ainsi, pour ces trois cancers, la taille moyenne des banques doit être au moins de 5 millions de lectures, et les prédictions pourraient potentiellement être améliorées en augmentant la profondeur de séquençage.

Ensuite, les prédictions obtenues sont saturées en fonction du nombre de patients pour l'ensemble des 11 cancers étudiés et à la fois pour les données RNA-seq d'ARNm et de miARN lorsque l'on utilise ce modèle de Cox multivarié.

Finalement, nous avons montré que la diminution des capacités prédictives induite par une diminution de la profondeur de séquençage n'est pas due au nombre plus faible de gènes détectés lorsque la taille des banques est plus faible, mais à la diminution du rapport signal sur bruit des données de comptage RNA-seq. Cette diminution induit une dégradation de la qualité des données et donc de la prédiction.

4.9 Références

ANDOR, N. et collab.. 2016, «Pan-cancer analysis of the extent and consequences of intratumor heterogeneity», *Nature Medicine*, vol. 22, nº 1, doi :10.1038/nm.3984, p. 105–113, ISSN 1546-170X. URL https://www.nature.com/articles/nm.3984, number : 1 Publisher : Nature Publishing Group. 113

BARTEL, D. P. 2018, «Metazoan MicroRNAs», *Cell*, vol. 173, nº 1, doi :10.1016/j.cell.2018.03.006, p. 20–51, ISSN 00928674. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418302861. 112

BASS, A. J. et collab.. 2019, «Determining sufficient sequencing depth in RNA-Seq differential expression studies», *bioRxiv*, doi :10.1101/635623. URL https://www.biorxiv.org/content/early/2019/05/13/635623. 113, 124

BENJAMINI, Y. et Y. HOCHBERG. 1995, «Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing», *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, nº 1, p. 289–300, ISSN 0035-9246. URL https://www.jstor.org/stable/2346101, publisher : [Royal Statistical Society, Wiley]. 115, 119, 131

BENJAMINI, Y. et T. P. SPEED. 2012, «Summarizing and correcting the GC content bias in high-throughput sequencing», *Nucleic Acids Research*, vol. 40, nº 10, doi : 10.1093/nar/gks001, p. e72–e72, ISSN 0305-1048. URL https://academic.oup.com/nar/article/40/10/e72/2411059, publisher : Oxford Academic. 116

CALIN, G. A. et C. M. CROCE. 2006, «MicroRNA signatures in human cancers», *Nat. Rev. Cancer*, vol. 6, nº 11, p. 857–866. 112

CHU, A., G. ROBERTSON, D. BROOKS, A. J. MUNGALL, I. BIROL, R. COOPE, Y. MA, S. JONES et M. A. MARRA. 2016, «Large-scale profiling of microRNAs for The Cancer Genome Atlas», *Nucleic Acids Research*, vol. 44, nº 1, doi :10.1093/nar/gkv808, p. e3, ISSN 0305-1048. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4705681/. 114

CIEŚLIK, M. et A. M. CHINNAIYAN. 2018, «Cancer transcriptome profiling at the juncture of clinical translation», *Nature Reviews Genetics*, vol. 19, nº 2, doi :10.1038/nrg.2017.96,

p. 93–109, ISSN 1471-0064. URL https://www.nature.com/articles/nrg.2017.96, number : 2 Publisher : Nature Publishing Group. 113

GERLINGER, M. et collab.. 2012, «Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing», *New England Journal of Medicine*, vol. 366, nº 10, doi :10.1056/NEJMoa1113205, p. 883–892, ISSN 0028-4793. URL https://doi.org/10.1056/NEJMoa1113205, publisher : Massachusetts Medical Society _eprint : https ://doi.org/10.1056/NEJMoa1113205. 113, 129

HEIMBERG, G. et collab.. 2016, «Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing», *Cell Systems*, vol. 2, nº 4, doi :10.1016/j.cels.2016.04.001, p. 239–250, ISSN 24054712. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216301090. 114

KLIEBENSTEIN, D. J. 2012, «Exploring the Shallow End ; Estimating Information Content in Transcriptomics Studies», *Frontiers in Plant Science*, vol. 3, doi :10.3389/fpls.2012.00213, ISSN 1664-462X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3437520/. 114

KUMAR-SINHA, C. et A. M. CHINNAIYAN. 2018, «Precision oncology in the age of integrative genomics», *Nature Biotechnology*, vol. 36, nº 1, doi :10.1038/nbt.4017, p. 46–60, ISSN 1546-1696. URL https://www.nature.com/articles/nbt.4017, number : 1 Publisher : Nature Publishing Group. 113

KURTZ, D. M. et collab.. 2019, «Dynamic Risk Profiling Using Serial Tumor Biomarkers for Personalized Outcome Prediction», *Cell*, vol. 178, nº 3, doi :10.1016/j.cell.2019.06.011, p. 699–713.e19, ISSN 1097-4172. 113

LIU, Y. et collab.. 2014, «RNA-seq differential expression studies : more sequence or more replication ?», *Bioinformatics*, vol. 30, nº 3, doi :10.1093/bioinformatics/btt688, p. 301–304, ISSN 1367-4803. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3904521/. 113

LU, J. et collab.. 2018, «Clinical significance of prognostic score based on age, tumor size, and grade in gastric cancer after gastrectomy», *Cancer Management and Research*, vol. 10, doi :10.2147/CMAR.S171663, p. 4279–4286, ISSN 1179-1322. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6183590/. 121

MAMMEDOV, T. et collab.. 2008, «A Fundamental Study of the PCR Amplification of GC-Rich DNA Templates», *Computational biology and chemistry*, vol. 32, nº 6, doi : 10.1016/j.compbiolchem.2008.07.021, p. 452–457, ISSN 1476-9271. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2727727/. 116

MATHÉ, E. A. et collab.. 2009, «MicroRNA Expression in Squamous Cell Carcinoma and Adenocarcinoma of the Esophagus : Associations with Survival», *Clinical Cancer Research*, vol. 15, nº 19, doi :10.1158/1078-0432.CCR-09-1467, p. 6192–6200, ISSN 1078-0432, 1557-3265. URL https://clincancerres.aacrjournals.org/content/15/19/6192, publisher : American Association for Cancer Research Section : Imaging, Diagnosis, Prognosis. 119

MCGRANAHAN, N. et C. SWANTON. 2017, «Clonal Heterogeneity and Tumor Evolution : Past, Present, and the Future», *Cell*, vol. 168, nº 4, doi :10.1016/j.cell.2017.01.018, p. 613–628, ISSN 1097-4172. 113, 129

MICHIELS, S. et collab.. 2016, «Statistical controversies in clinical research : prognostic gene signatures are not (yet) useful in clinical practice», *Annals of Oncology : Official Journal of the European Society for Medical Oncology*, vol. 27, nº 12, doi :10.1093/annonc/mdw307, p. 2160–2167, ISSN 1569-8041. 113

MILANEZ-ALMEIDA, P. et collab.. 2020, «Cancer prognosis with shallow tumor RNA sequencing», *Nature Medicine*, vol. 26, nº 2, doi :10.1038/s41591-019-0729-3, p. 188–192, ISSN 1078-8956, 1546-170X. URL http://www.nature.com/articles/s41591-019-0729-3. 114, 121, 122, 128, 133

MORTAZAVI, A. et collab.. 2008, «Mapping and quantifying mammalian transcriptomes by RNA-Seq», *Nature Methods*, vol. 5, nº 7, doi :10.1038/nmeth.1226, p. 621–628, ISSN 1548-7091, 1548-7105. URL http://www.nature.com/articles/nmeth.1226. 113, 129

PAWITAN, Y. et collab.. 2005, «False discovery rate, sensitivity and sample size for microarray studies», *Bioinformatics*, vol. 21, nº 13, doi :10.1093/bioinformatics/bti448, p. 3017–3024, ISSN 1367-4803, 1460-2059. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti448. 113

RAPAPORT, F., R. KHANIN, Y. LIANG, M. PIRUN, A. KREK, P. ZUMBO, C. E. MASON, N. D. SOCCI et D. BETEL. 2013, «Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data», *Genome Biology*, vol. 14, nº 9, doi :10.1186/gb-2013-14-9-r95, p. R95, ISSN 1465-6906. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-9-r95. 113

RICKETTS, C. J. et collab.. 2018, «The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma», *Cell Reports*, vol. 23, nº 1, doi :10.1016/j.celrep.2018.03.075, p. 313–326.e5, ISSN 2211-1247. URL https://www.cell.com/cell-reports/abstract/S2211-1247(18)30436-4, publisher : Elsevier. 122

RISSO, D. et collab.. 2011, «GC-Content Normalization for RNA-Seq Data», *BMC Bioinformatics*, vol. 12, nº 1, doi :10.1186/1471-2105-12-480, p. 480, ISSN 1471-2105. URL https://doi.org/10.1186/1471-2105-12-480. 116

ROBINSON, D. G. et J. D. STOREY. 2014, «subSeq : Determining Appropriate Sequencing Depth Through Efficient Read Subsampling», *Bioinformatics*, vol. 30, nº 23, doi :10.1093/bioinformatics/btu552, p. 3424–3426, ISSN 1460-2059, 1367-4803. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu552. 123

ROSENBERG, J. et collab.. 2005, «The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database», *Breast Cancer Research and Treatment*, vol. 89, nº 1, doi :10.1007/s10549-004-1470-1, p. 47–54, ISSN 0167-6806. 121

RUPAIMOOLE, R. et F. J. SLACK. 2017, «MicroRNA therapeutics : towards a new era for the management of cancer and other diseases», *Nature Reviews. Drug Discovery*, vol. 16, nº 3, doi :10.1038/nrd.2016.246, p. 203–222, ISSN 1474-1784. 112

SENFT, D. et collab.. 2017, «Precision Oncology : The Road Ahead», *Trends in Molecular Medicine*, vol. 23, nº 10, doi :10.1016/j.molmed.2017.08.003, p. 874–898, ISSN 1471-4914. URL http://www.sciencedirect.com/science/article/pii/S1471491417301430. 113

SHI, X.-H. et collab.. 2018, «A Five-microRNA Signature for Survival Prognosis in Pancreatic Adenocarcinoma based on TCGA Data», *Scientific Reports*, vol. 8, nº 1, doi : 10.1038/s41598-018-22493-5, p. 7638, ISSN 2045-2322. URL https://www.nature.com/articles/s41598-018-22493-5. 112

SIMS, D. et collab.. 2014, «Sequencing depth and coverage : key considerations in genomic analyses», *Nature Reviews Genetics*, vol. 15, nº 2, doi :10.1038/nrg3642, p. 121–132, ISSN 1471-0056, 1471-0064. URL http://www.nature.com/articles/nrg3642. 113, 129

TARAZONA, S., F. GARCIA-ALCALDE, J. DOPAZO, A. FERRER et A. CONESA. 2011, «Differential expression in RNA-seq : A matter of depth», *Genome Research*, vol. 21, nº 12, doi : 10.1101/gr.124321.111, p. 2213–2223, ISSN 1088-9051. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.124321.111. 113, 124

TERNÈS, N. et collab.. 2017, «Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces», *Biometrical Journal. Biometrische Zeitschrift*, vol. 59, nº 4, doi :10.1002/bimj.201500234, p. 685–701, ISSN 1521-4036. 122

VERWEIJ, P. J. M. et H. C. VAN HOUWELINGEN. 1994, «Penalized likelihood in cox regression», *Statistics in Medicine*, vol. 13, nº 23-24, doi :10.1002/sim.4780132307, p. 2427–2436. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780132307. 114

WEI, H. et collab.. 2017, «MiR-638 inhibits cervical cancer metastasis through Wnt/-catenin signaling pathway and correlates with prognosis of cervical cancer patients», *European Review for Medical and Pharmacological Sciences*, vol. 21, nº 24, doi :10.26355/eurrev_201712_13999, p. 5587–5593, ISSN 2284-0729. 122

YANG, H., H. SU, N. HU, C. WANG, L. WANG, C. GIFFEN, A. M. GOLDSTEIN, M. P. LEE et P. R. TAYLOR. 2020, «Integrated analysis of genome-wide miRNAs and targeted gene expression in esophageal squamous cell carcinoma (ESCC) and relation to prognosis», *BMC Cancer*, vol. 20, nº 1, doi :10.1186/s12885-020-06901-6, p. 388, ISSN 1471-2407. URL https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-06901-6. 119

YU, S.-L. et collab.. 2008, «MicroRNA signature predicts survival and relapse in lung cancer», *Cancer Cell*, vol. 13, nº 1, doi :10.1016/j.ccr.2007.12.008, p. 48–57, ISSN 1535-6108. 112

ZHANG, J. et collab.. 2015, «A Seven-microRNA Expression Signature Predicts Survival in Hepatocellular Carcinoma», *PLoS ONE*, vol. 10, nº 6, doi :10.1371/journal.pone.0128628, ISSN 1932-6203. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4457814/. 112

ZHOU, X. et collab.. 2016, «A panel of 13-miRNA signature as a potential biomarker for predicting survival in pancreatic cancer», *Oncotarget*, vol. 7, nº 43, doi :10.18632/oncotarget.11903, p. 69 616–69 624, ISSN 1949-2553. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5342502/. 112

ZHUO, M. et collab.. 2019, «Survival analysis via nomogram of surgical patients with malignant pleural mesothelioma in the Surveillance, Epidemiology, and End Results database», *Thoracic Cancer*, vol. 10, nº 5, doi :10.1111/1759-7714.13063, p. 1193–1202, ISSN 1759-7706. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6501014/. 122

Chapitre 5

Conclusions et perspectives

Sommaire

5.1 Conclusions générales 140

5.2 Résultats préliminaires et perspectives 141

5.2.1 Utilisation des termes d'interaction entre gènes 141

5.2.2 Ajout de nouvelles données cliniques pour prédire la survie 143

5.2.3 Approches « gros grains » 145

5.2.4 Proportions de types cellulaires pour prédire la survie 145

5.2.5 Prise en compte de l'hétérogénéité intra-tumorale dans la prédiction de la survie à partir de données mRNA-seq 147

5.2.6 Utilisation d'autres algorithmes de prédiction de survie 148

5.2.7 Intégration de données « multi-omiques » pour prédire la survie . . 149

5.3 Conclusions pratiques 150

5.4 Références 151

5.1 Conclusions générales

Pour rappel, les objectifs de cette thèse étaient :

1. de comparer les différentes formes classiques de pénalisation du modèle de Cox (*i.e.* ridge, lasso, elastic net, adaptive elastic net) sur les données réelles de TCGA et données simulées (chapitre 2).
2. d'étudier l'impact du pré-filtrage univarié des gènes sur les prédictions de la survie obtenues avec le modèle de Cox pénalisé (chapitre 3).
3. d'étudier l'impact de la profondeur de séquençage sur les prédictions obtenues avec le modèle de Cox pénalisé (chapitre 4).

Tout d'abord, nous avons montré que les prédictions obtenues avec $\lambda_{\min}$ étaient en moyenne meilleures que celle obtenues avec λ_{1se} (partie 2.2.1). Ensuite, la pénalisation ridge obtient en moyenne de meilleurs résultats que lasso, elastic net, et adaptive elastic net (partie 2.2.3). Cependant, la pénalisation ridge assigne des coefficients de régression $\beta_i, i = 1, \ldots, p$, non nul à chaque gène, et ne permet pas de sélectionner un sous-ensemble de prédicteurs. Parmi les méthodes qui permettent d'obtenir un modèle parcimonieux (*i.e.* lasso, elastic net, adaptive elastic net), la pénalisation elastic net obtient en moyenne les meilleurs résultats. Ces premières conclusions sont à nuancer par le fait que les résultats dépendent fortement des jeux de données : la qualité des prédictions obtenues dépend plus du type de cancer que de la méthode utilisée. De plus, le C-index et la p-valeur du modèle de Cox sont corrélés entre eux et sensibles à la méthode choisie, alors que l'IBS est moins corrélé aux autres métriques, et moins sensible à la méthode étudiée.

Ensuite, nous avons montré que les données mRNA-seq couplées à des variables cliniques classiques permettent d'obtenir de meilleures prédictions que ces mêmes variables cliniques seules pour 12 des 26 cancers étudiés dans le chapitre 2 (partie 2.3). De même, nous avons observé que l'ajout des données miRNA-seq aux données cliniques classiques permet d'améliorer les prédictions pour 7 des 11 cancers étudiés dans le chapitre 4 (partie 4.4).

La mise en place d'une procédure de simulation autorisant le contrôle du niveau de corrélation entre les données d'expression génétique et la survie (partie 2.4.1) nous a permis de montrer que les performances de prédiction obtenues par l'Oracle (*i.e.* estimateurs du C-index, de l'IBS, et de la p-valeur du modèle de Cox à partir du vecteur $\boldsymbol{\beta}$ utilisé pour les simulations) sont loin d'être atteintes par le modèle de Cox pénalisé (partie 2.5). Ensuite, les performances de sélection des trois méthodes de pénalisation parcimonieuses (*i.e.* lasso, elastic net, adaptive elastic net) sont mauvaises, à la fois en terme de stabilité, de sensibilité, et de taux de fausses découvertes (partie 2.6). Nous avons alors montré que ces mauvaises performances de sélection ont un impact négatif sur la qualité de la prédiction (partie 2.7). Identifier les gènes corrélés à la survie apparaît donc important dans l'optique d'améliorer significativement les prédictions obtenues avec les données RNA-

seq. Dans le but d'accroître les performances de sélection (et donc de prédiction), nous nous sommes intéressés à des méthodes de pré-filtrage univarié comme première étape de sélection.

Ainsi, dans le chapitre 3, nous avons tout d'abord montré que le pré-filtrage bi-dimensionnel sur les p-valeurs du modèle de Cox univarié et les écarts interquartiles permet d'améliorer les prédictions, mais seulement marginalement en valeur absolue (partie 3.2). En effet, l'augmentation moyenne du C-index pour l'ensemble des 16 cancers étudiés dans ce chapitre n'est que de 0,016 pour la pénalisation lasso. Cependant, le pré-filtrage univarié des gènes permet de réduire simplement et d'un facteur 20 en moyenne le nombre de prédicteurs à un ensemble de gènes a priori pertinents pour prédire la survie. Cette réduction de la dimension permet d'augmenter la stabilité des gènes restants pour composer les scores de risques (indices pronostiques). Enfin, le pré-filtrage sur l'écart-interquartile des gènes semble plus efficace que celui sur la p-valeur du modèle de Cox univarié, et les prédictions obtenues avec le pré-filtrage bi-dimensionnel sont meilleures que celles obtenues avec l'algorithme ISIS (partie 3.4).

Bien que les coûts de la technologie RNA-seq aient diminué drastiquement au cours de la dernière décennie, ils restent trop élevés pour que le séquençage soit utilisé en clinique et pour de grandes cohortes de patients. Ainsi, afin d'optimiser le nombre d'échantillons séquencés et les prédictions sous contrainte de coûts, Milanez-Almeida et collab. [2020] ont montré qu'il était possible de diminuer d'un facteur 100 la profondeur de séquençage des données mRNA-seq. Dans le chapitre 4, nous avons confirmé ce résultat et nous nous sommes intéressés à l'impact de la profondeur de séquençage des données miRNA-seq sur les prédictions. Nous avons montré que les résultats sont plus variables pour les données miRNA-seq que pour les données mRNA-seq. En effet, la taille des banques peut être diminuée d'un facteur 1000 pour UVM, alors qu'elle ne peut pas être diminuée pour MESO, CESC, ou LIHC.

Dans les prochains paragraphes et à partir des conclusions établies ci-dessus, nous dresserons quelques perspectives de l'utilisation de données génomiques pour l'aide au pronostic de patients atteints de cancer.

5.2 Résultats préliminaires et perspectives

5.2.1 Utilisation des termes d'interaction entre gènes

L'existence de processus d'inhibition et d'activation entre gènes a montré un intérêt pronostique dans le cancer [Baldus et collab., 2004]. Ces interactions entre les gènes peuvent être prise en compte dans les modèles de survie [Ashworth et collab., 2011; Magen et collab., 2019]. Pour mieux prendre en compte ces effets jumelés sur la survie, nous avons ajouté les termes quadratiques des gènes sélectionnés par la pénalisation elastic net dans le modèle. La fonction de risque instantané s'écrit alors [Wu et collab., 2018] :

$$\lambda(t;X_1,...,X_p) = \lambda_0(t)\exp(f(X_1,...,X_p)),$$

avec :

— $f(X_1,...,X_p) = \beta_1 X_{(1)} + ... + \beta_q X_{(q)} + \beta_{12} X_{(1)}X_{(2)} + ... + \beta_{q-1,q} X_{(q-1)}X_{(q)}$.

— q le nombre de gènes sélectionnés par la pénalisation elastic net.

Les nouveaux prédicteurs se définissent donc comme les niveaux d'expression des gènes sélectionnés par la pénalisation elastic net (termes « linéaires »), et les termes d'interaction $X_{(i)}X_{(j)}$ de ces gènes (termes « quadratiques »). Nous apprenons alors un deuxième modèle de Cox avec pénalisation elastic net sur ces nouveaux prédicteurs, et nous calculons les différentes mesures d'évaluation des prédictions (*i.e.* C-index, p-valeur du modèle de Cox univarié, IBS). Chaque gène qui compose les termes d'interaction a été sélectionné dans la première étape (*i.e.* modèle de Cox avec pénalisation elastic net sur le niveau d'expression des gènes), et cette approche est dite « à forte hérédité » [Bien et collab., 2013; Lim et Hastie, 2015; Wang et collab., 2014; Wu et collab., 2018].

Nous observons que l'ajout des termes d'interactions dans le modèle de Cox pénalisé ne permet pas d'améliorer les prédictions (Fig. 5.1.A, données non montrées pour l'IBS et la p-valeur du modèle de Cox univarié). Cependant, parmi les co-variables sélectionnées dans le modèle final par elastic net, presque uniquement des termes d'interaction sont sélectionnés (Fig. 5.1.B). Mieux comprendre ce phénomène et prendre en compte ces interactions semblent des perspectives intéressantes à envisager.

D'autres modèles, comme les méthodes à noyaux [de Vlaming et Groenen, 2015] ou les réseaux de neurones profonds [Huang et collab., 2019] peuvent être envisagés pour prendre en compte les interactions entre les gènes et les effets non linéaires sur la survie. Cependant, une meilleure prise en compte de la connaissance biologique dans les modèles mathématiques pourrait permettre d'accroître leur pertinence et leur pouvoir prédictif [Domany, 2014].

Par exemple, une voie de signalisation se définit comme l'ensemble des évènements intracellulaires déclenchés par un signal à l'un des récepteurs de la cellule, jusqu'à la réponse cellulaire. Ces évènements sont contrôlés par des protéines de signalisation intracellulaires, et donc par l'expression des gènes. Ainsi, une voie de signalisation peut se modéliser par un graphe qui relie les gènes qui la composent entre eux. Différentes bases de données, comme `KEGG` [Kanehisa et Goto, 2000; Kanehisa et collab., 2019] ou `MSigDB` [Liberzon et collab., 2011; Subramanian et collab., 2005] décrivent les liens qui existent entre les gènes. Cette information biologique peut être prise en compte dans les modèles, et de nombreuses approches existent [Dereli et collab., 2019; Fa et collab., 2019; Zheng et collab., 2020].

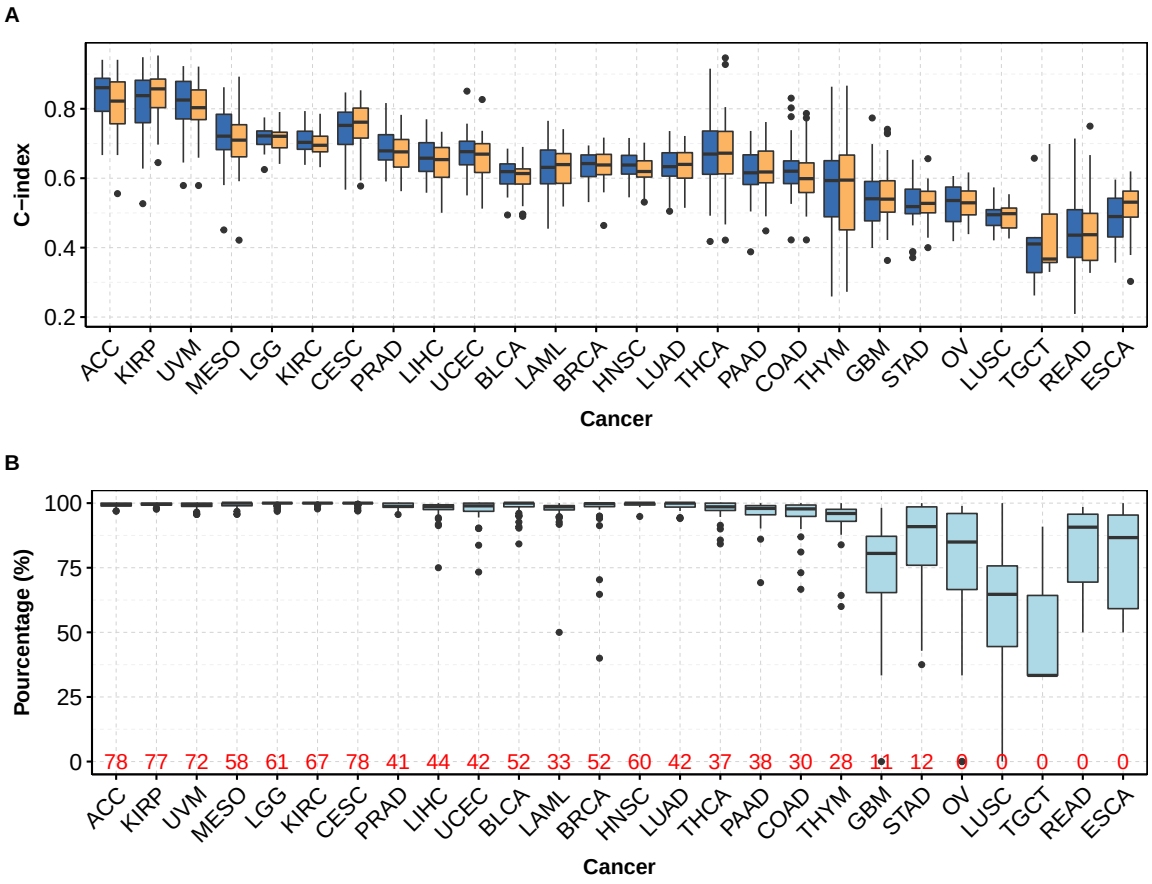


FIGURE 5.1 – **C-index obtenus avec elastic net avec uniquement le niveau d'expression des gènes (bleu), et avec l'ajout des termes d'interactions (jaune) (A), et proportions de termes quadratiques sélectionnés.**

Nous calculons les C-index par 10 répétitions du validation croisée (K=5). Les termes d'interactions sont calculés à partir des gènes sélectionnés par la pénalisation elastic net.

5.2.2 Ajout de nouvelles données cliniques pour prédire la survie

Les analyses pan-cancer que nous avons menés dans ce manuscrit permettent de tester une méthodologie ou de comparer des méthodes sur un grand nombre de jeux de données. Cependant, il devient difficile de se focaliser sur les particularités des cancers étudiés. Par exemple, le statut des œstrogènes (négatif, intermédiaire, ou positif) est un marqueur bien connu et utilisé en pratique comme marqueur prédictif de la survie dans le cancer du sein [VOLKMANN et collab., 2019]. Ainsi, en ajoutant ce statut aux variables cliniques utilisées jusqu'à présent dans ce manuscrit pour BRCA (*i.e.* âge, genre, « T », « N », « M »), les prédictions obtenues uniquement avec les données cliniques sont significativement améliorées (p-valeur < 0.001, test des rangs signés de Wilcoxon unilatéral, Fig. 5.2.A).

Ensuite, nous avons identifié des variables cliniques reliées à la survie des patients en calculant les p-valeurs de modèles de Cox univarié pour chacune d'entre elles pour KIRC (variables qualitatives et quantitatives). Après correction par la méthode de Benjamini-Hochberg, nous avons identifié deux nouvelles variables intéressantes pour prédire la

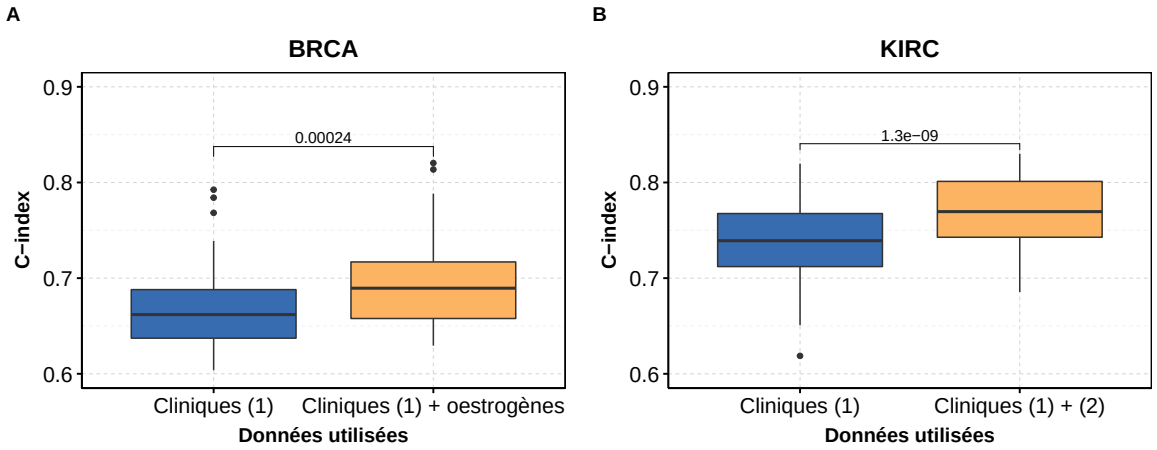


FIGURE 5.2 – **C-index obtenus après l'ajout de nouvelles variables cliniques pour BRCA (A) et KIRC (B).**

Les boites bleues correspondent aux C-index obtenus avec les variables cliniques utilisées dans les chapitres précédents (âge, genre, T, N, M pour BRCA (A) ; âge, genre, T, N, M, grade pour KIRC (B)). Les boites oranges correspondent aux C-index obtenus après ajout du statut des récepteur d'œstrogènes pour BRCA (A), et du pourcentage de nécrose et de noyaux de cellules tumorales pour KIRC (B).

Nous calculons les C-index par 10 répétitions d'une validation croisée (K=5). Les p-valeurs de tests des rangs signés de Wilcoxon unilatéraux permettant de tester si les C-index obtenus après ajout des nouvelles variables cliniques ont augmenté sont indiquées sur chaque graphique.

survie :

— le pourcentage de nécrose (*i.e.* pourcentage de cellules mortes dans l'échantillon).

— le pourcentage de cellules tumorales dans l'échantillon.

Après l'ajout de ces deux variables dans le modèle de Cox multivarié, les C-index calculés sont significativement supérieurs à ceux obtenus avec les variables cliniques utilisées dans les chapitres précédents pour ce cancer (*i.e.* âge, genre, « T », « N », « M », grade) (p-valeur < 0.001, test des rangs signés de Wilcoxon unilatéral, Fig. 5.2.B).

De telles analyses basées à la fois sur des connaissances cliniques de la maladie (*e.g.* statut des œstrogènes pour le cancer du sein) et sur une analyse statistique permettant d'identifier des variables potentiellement intéressantes (*e.g.* p-valeur du modèle de Cox univarié sur les données cliniques de TCGA) devraient être menées individuellement pour chaque cancer [HERRMANN et collab., 2020]. Cela permettrait d'améliorer les scores de risque associés à la survie, et donc d'aider à la prise en charge clinique des patients. Ainsi, travailler à l'élaboration de scores de risque basés sur des variables cliniques classiques et spécifiques du cancer étudié en les comparant à ceux utilisés par les cliniciens pourrait aboutir à l'amélioration rapide du pronostic des patients, et donc de leur suivi.

5.2.3 Approches « gros grains »

Pour remédier à la « malédiction de la grande dimension », des approches dites « gros grains » ont été proposées [Drier et collab., 2013; Hänzelmann et collab., 2013; Vaske et collab., 2010]. L'idée de cette démarche est d'assigner un score à des sous-groupes de gènes (*e.g.* gènes d'une même voie de signalisation), et d'utiliser ces scores dans un modèle de Cox pour prédire la survie, identifier des sous-types de cancers, ou des processus biologiques dérégulés dans le cancer [Domany, 2014]. Cette stratégie a montré des résultats de prédiction intéressants en pratique [Fa et collab., 2019].

Zheng et collab. [2020] ont comparé les prédictions (C-index) obtenues en utilisant les niveaux d'expression des gènes et les scores assignés aux voies de signalisation. Les voies de signalisation utilisées proviennent de la base de données MSigDB [Subramanian et collab., 2005], et la méthode GSVA [Hänzelmann et collab., 2013] permet d'assigner un score à chaque voie de signalisation et à chaque patient. Les auteurs ont montré, sur données réelles et simulées, que les performances de prédiction (C-index) obtenues avec les scores assignés aux voies de signalisation sont équivalentes à celles obtenues avec le niveau d'expression des gènes lorsqu'il y a présence de fortes corrélations, et meilleures lorsque ces corrélations sont faibles.

Ces approches semblent donc intéressantes à considérer pour travailler indépendamment sur des groupes de gènes restreints et donc réduire la dimension, garder une interprétation biologique, et augmenter la robustesse et les capacités de prédiction. L'estimation des types cellulaires présents dans un échantillon est un domaine de recherche en plein développement [Cobos et collab., 2020], et qui possède tous les avantages que décrits dans ce paragraphe. Nous étudierons leur pouvoir prédictif dans la prochaine partie.

5.2.4 Proportions de types cellulaires pour prédire la survie

L'estimation de la composition cellulaire des tissus tumoraux (qualifié ici d'« hétérogénéité intra-tumorale ») a de nombreuses applications pratiques. Par exemple, tester si un gène est différentiellement exprimé peut être biaisée si le gène étudié est spécifique d'un type cellulaire présent dans des proportions différentes suivant les patients [Aran et collab., 2015]. Ensuite, l'analyse des cellules immunitaires présentes dans la tumeur demeure essentielle pour une meilleure compréhension biologique du cancer. Cela ouvre la voie à des traitements ciblés sur des types cellulaires précis à travers l'immunothérapie [Hendry et collab., 2017; Sharma et collab., 2019]. Enfin, ajuster le niveau d'expression des gènes en prenant en compte l'hétérogénéité intra-tumorale permet d'obtenir des classifications plus pertinentes [Aran et collab., 2015; Elloumi et collab., 2011]. Les sous-types obtenus ont alors plus de sens biologique.

Dans cette partie, nous étudierons les capacités prédictives des données d'hétérogénéité intra-tumorale. Nous utiliserons deux méthodes classiques qui ont obtenu de bons résultats dans des études comparatives indépendantes [Avila Cobos et collab., 2018;

STURM et collab., 2019] : CIBERSORT [NEWMAN et collab., 2015] et xCell [ARAN et collab., 2017]. CIBERSORT permet une estimation de la proportion de 22 types de cellules immunitaires, et xCell de 64 types cellulaires à la fois immunitaires et non immunitaires.

Enfin, la « pureté de la tumeur » se définit comme le pourcentage de cellules tumorales dans l'échantillon, et demeure intéressant à utiliser en pratique, notamment pour la prédiction [MAO et collab., 2018]. Nous utiliserons l'estimation de la pureté des tumeurs de TCGA faites par ARAN et collab. [2015] dans un modèle de Cox univarié pour prédire la survie. Parmi les 26 cancers étudiés, les données CIBERSORT, xCell, et de pureté de la tumeur sont disponibles pour 18 cancers.

Nous observons que le C-index médian obtenu avec les données mRNA-seq est supérieur à ceux obtenus avec les données de proportions de types cellulaires (*i.e.* CIBERSORT et xCell) et de pureté de la tumeur (Fig. 5.3). Les deux seules exceptions sont pour UCEC et READ, pour lesquels le C-index médian obtenu avec CIBERSORT est plus important, mais avec une grande variance. La variance des C-index obtenus avec les données CIBERSORT, xCell, et de pureté de la tumeur est en effet plus importante pour l'ensemble des cancers étudiés. Nous tirons les mêmes conclusions avec la p-valeur du modèle de Cox univarié (*i.e.* les prédictions obtenues avec les données mRNA-seq sont meilleures), et les résultats sont équivalents pour l'IBS.

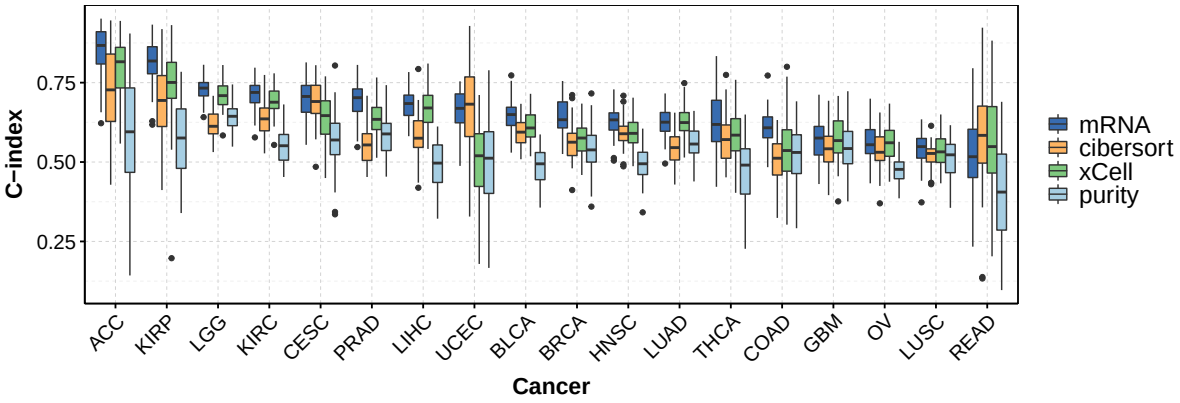


FIGURE 5.3 – **C-index obtenus avec les données mRNA-seq, de proportions de types cellulaires (*i.e.* CIBERSORT et xCell), et de pureté de la tumeur pour 18 cancers de TCGA.**
Nous calculons les C-index par 10 répétitions du validation croisée (K=5).

De plus, parmi les trois méthodes testées d'estimation de l'hétérogénéité, xCell obtient globalement les meilleurs résultats. Les deux seules exceptions sont pour CESC et UCEC pour lesquels les C-index sont significativement supérieurs pour CIBERSORT (p-valeur < 0,01, test des rangs signés de Wilcoxon). Nous observons les mêmes résultats pour la p-valeur du modèle de Cox univarié, et les résultats sont équivalents pour l'IBS.

Ainsi, ces données d'estimations des populations cellulaires ne permettent pas d'atteindre de meilleures prédictions que les données mRNA-seq. Dans la partie suivante, nous allons essayer d'améliorer les prédictions en mixant les indices pronostiques obte-

nus avec ces différents types de données (*i.e.* mRNA-seq, proportions de types cellulaires, pureté de la tumeur), et ceux obtenus avec les données cliniques.

5.2.5 Prise en compte de l'hétérogénéité intra-tumorale dans la prédiction de la survie à partir de données mRNA-seq

Une approche intuitive pour prendre en compte l'hétérogénéité intra-tumorale dans la prédiction de la survie à partir des données RNA-seq consiste à utiliser un modèle de Cox multivarié composé des variables cliniques classiques, et des indices pronostiques calculés séparément à partir des données mRNA-seq, xCell, et de pureté de la tumeur. En revanche, cette méthode n'apporte pas de pouvoir prédictif au modèle (Fig. 5.4).

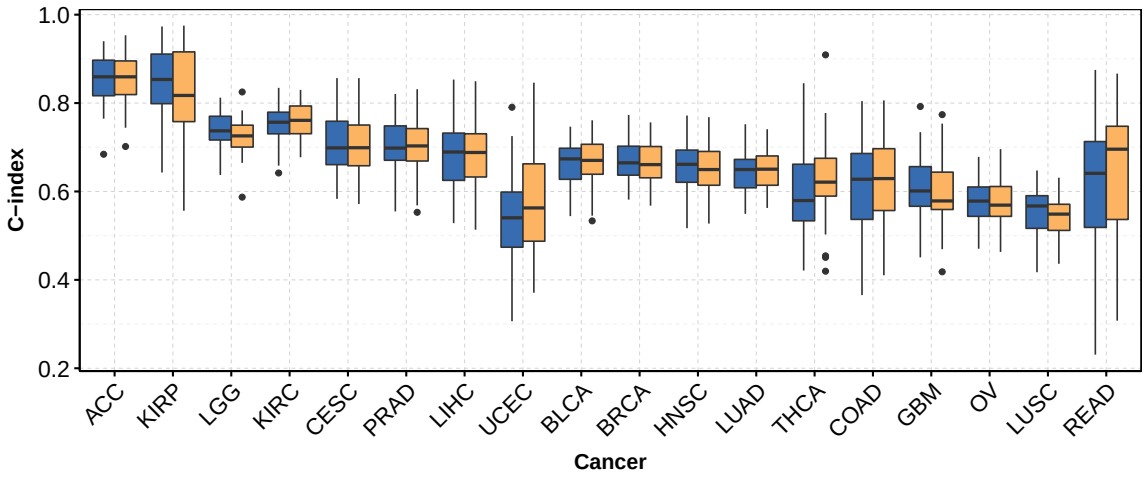


FIGURE 5.4 – **C-index obtenus avec les variables cliniques et les indices pronostiques des données mRNA-seq (bleu), et avec les variables cliniques et les indices pronostiques des données mRNA-seq, xCell, et de pureté de la tumeur (jaune).**
Nous avons calculé les C-index par 10 répétitions d'une validation croisée (K=5).

Une autre méthodologie pour prendre en compte l'hétérogénéité intra-tumorale peut être de séparer les patients en plusieurs sous-groupes, et d'apprendre un modèle de Cox pour chaque sous-groupe, indépendamment. Plus généralement, l'existence de groupes de patients possédant des caractéristiques génomiques, cellulaires, ou cliniques distinctes peut engendrer une trop grande hétérogénéité au sein des jeux de données et détériorer les résultats obtenus sur les prédictions. Séparer les patients suivant leurs caractéristiques génomiques, cellulaires, ou cliniques pour les analyser séparément peut ainsi être une solution à envisager.

Il est important de noter que les données de pureté de la tumeur doivent être considérées pour stratifier les patients en différents groupes à partir des données d'expression génétique. En effet, [ARAN et collab., 2015] ont montré que les sous-types de cancer obtenus sont en fait des groupes de patients possédant des puretés de tumeur similaires, et que le *clustering* regroupe des gènes dont les corrélations avec la pureté de la tumeur sont

similaires (Fig. 5.5).

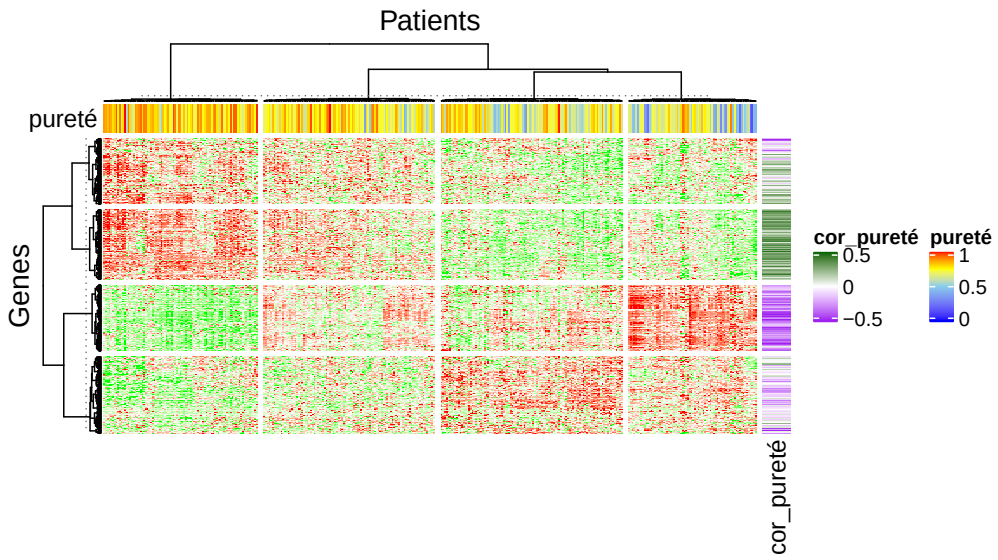


FIGURE 5.5 – ***Heatmap*** **du niveau d'expression standardisé des gènes pour BLCA, et corrélation avec la pureté de la tumeur.**
Nous avons retenu les 2000 gènes qui ont les plus faibles p-valeurs d'un modèle de Cox univarié. La corrélation des gènes avec la pureté de la tumeur est indiquée à droite (vert et violet) pour chaque gène, et la pureté de tumeur est indiquée en haut du graphique pour chaque patient (bleu, jaune, et rouge).
Au sein de la *heatmap*, le rouge correspond à un niveau d'expression élevé, et le vert à un niveau d'expression faible.

5.2.6 Utilisation d'autres algorithmes de prédiction de survie

Dans ce manuscrit, nous nous sommes focalisés sur le modèle de Cox multivarié. D'autres algorithmes existent, et il serait intéressant de les comparer sur différents types de données (*i.e.* différents cancers, données réelles et simulées). Par exemple, nous avons appliqué un algorithme de « forêts aléatoires » [PROBST et collab., 2018] sur les données miRNA-seq que nous avons analysé au chapitre 4 et calculé 50 C-index par 10 répétitions d'une validation croisée (K=5). Ces méthodes ne reposent sur aucun modèle paramétrique et peuvent être plus flexibles que le modèle de Cox pour s'adapter à des données réelles. Néanmoins, les résultats obtenus sont équivalents ou moins bons que ceux obtenus avec le modèle de Cox pénalisé (ridge) (Fig. 5.6).

En plus des arbres aléatoires, d'autres approches existent comme la régression des moindres carrés partiels [BASTIEN et collab., 2015; CHUN et KELEŞ, 2010], et elles mériteraient d'être investiguées et comparées.

L'utilisation de données multi-omiques peut permettre une meilleure classification des patients en sous-types de cancer [DE MEULDER et collab., 2018]. Dans la prochaine partie, nous nous focaliserons sur l'intégration des données multi-omiques pour prédire la survie.

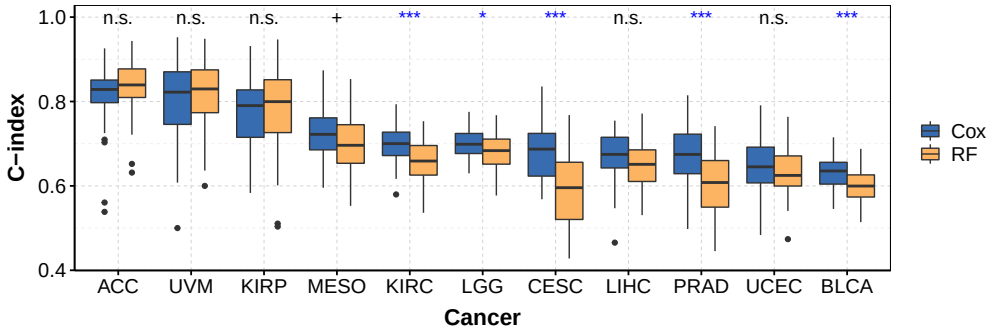


FIGURE 5.6 – **C-index obtenus avec le modèle de Cox pénalisé (ridge, bleu) et un algorithme de forêts aléatoires (jaune) sur les données miRNA-seq.**
Nous calculons les C-index par 10 répétitions du validation croisée (K=5), et nous indiquons le niveau de significativité d'un test de Wilcoxon en haut du graphique. n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

5.2.7 Intégration de données « multi-omiques » pour prédire la survie

Dans ce manuscrit, nous nous sommes intéressés au niveau d'expression des gènes (ARNm et miARN) grâce à la technologie RNA-seq (partie 1.3), mais d'autres types de données sont également mesurables. Par exemple, la méthylation de l'ADN consiste en l'ajout d'un groupement méthyle sur un nucléotide et permet l'activation ou la répression des gènes. Ce processus joue un rôle important dans le cancer [WANG et LEI, 2018]. De plus, la variabilité du nombre de copies d'un gène (*Copy Number Variation*, CNV) correspond à la présence de plusieurs copies d'un même gène dans le génome, et demeure aussi impliqué dans le cancer [ZHANG et collab., 2016].

Les analyses multi-omiques consistent en l'intégration de différents types de données omiques. La prédiction de la survie à partir de données multi-omiques demeure un pan actif de la recherche, et de nombreuses méthodes existent [WU et collab., 2019]. Différentes stratégies peuvent être adoptées pour l'intégration des différents types de données : les données peuvent être concaténées pour former une unique matrice (intégration « précoce »), ou analysées séparément puis assemblées (intégration « tardive ») [RAPPOPORT et SHAMIR, 2018].

Cependant, la valeur ajoutée des données multi-omiques par rapport aux données omiques analysées séparément ne fait pas consensus. Par exemple, ZHAO et collab. [2015] ont montré que parmi les données de méthylation, de CNV, d'ARNm et de miARN, les données d'ARNm associées aux données cliniques obtiennent des prédictions meilleures ou équivalentes à celles observées avec l'ensemble des types de données. Ce résultat a cependant été observé sur uniquement quatre cancers de TCGA, et avec une méthodologie précise. Ainsi, pour répondre à ces questions (*i.e.* intérêt des données multi-omiques pour prédire la survie et comparaison de méthodes), HERRMANN et collab. [2020] ont analysés 18 cancers de TCGA et 11 méthodes d'intégration de données. Les résultats diffèrent

suivant les jeux de données et la métrique utilisée (*i.e.* C-index, IBS), mais la méthode `blockForest` [HORNUNG et WRIGHT, 2019] obtient en moyenne les meilleurs résultats. Cependant, les prédictions obtenues avec cette méthode sont seulement très légèrement supérieures à celles obtenues avec uniquement les données cliniques.

D'autres modèles, comme les méthodes à noyaux ZHU et collab. [2017] ou les réseaux de neurones profonds [HUANG et collab., 2019] peuvent être envisagées pour l'intégration de données multi-omiques.

5.3 Conclusions pratiques

Finalement, nous recommandons l'utilisation du pré-filtrage bi-dimensionnel (*i.e.* sur la p-valeur du modèle de Cox univarié et l'écart-interquartile de chaque gène) puis de la pénalisation elastic net dans un modèle multivarié.

Ensuite, les données mRNA-seq peuvent aider aux pronostics des patients par rapport aux données cliniques classiques (parties 2.3 et 4.4). Cependant, les coûts sont encore élevés pour ue utilisation routinière (partie 4.1.3), et l'utilisation de variables cliniques spécifiques au cancer étudié pourrait permettre d'obtenir des prédictions équivalentes à celles obtenues avec les variables génomiques.

Ainsi, nous recommandons le développement de scores de risques basés sur des variables cliniques classiques et spécifiques du cancer étudié, et de les comparés aux scores classiquement utilisés en clinique. Dans cette optique, le modèle de Cox semble une approche intéressante car elle permet de prendre en compte des variables catégorielles et quantitatives et obtient de bons résultats en pratique. De plus, nous encourageons la prise en compte de connaissances cliniques et biologiques spécifiques de la maladie dans l'élaboration des modèles prédictifs. Dans cette optique, des collaborations entre biologistes, cliniciens, et bio-informaticiens sont essentiels et peuvent aboutir à de réels bénéfices pour les patients.

Les prédictions oracles sont loin d'être atteintes par le modèle de Cox pénalisé dans les simulations. De plus, il apparaît difficile d'évaluer si les prédictions maximales que l'on peut extraire des données réelles sont atteintes, ou si elles peuvent être améliorées par l'élaboration de nouvelles méthodes. Nous préconisons ainsi l'utilisation de simulations permettant de comparer les prédictions oracles aux prédictions obtenues par les modèles.

Enfin, bien que les coûts de la technologie RNA-seq aient diminué drastiquement au cours de la dernière décennie, ils restent élevés et demeurent une limitation importante à l'utilisation de données transcriptomiques en clinique et en recherche. Nous encourageons ainsi l'étude de l'impact de la réduction de la qualité des données (et donc des coûts) sur les prédictions pour un ensemble plus large de données génomiques et de méthodes. L'objectif est d'optimiser les prédictions et le nombre d'échantillons séquencés sous contrainte de coûts.

5.4 Références

ARAN, D. et collab.. 2015, «Systematic pan-cancer analysis of tumour purity», *Nature Communications*, vol. 6, nº 1, doi :10.1038/ncomms9971, p. 8971, ISSN 2041-1723. URL https://www.nature.com/articles/ncomms9971, number : 1 Publisher : Nature Publishing Group. 145, 146, 147

ARAN, D. et collab.. 2017, «xCell : digitally portraying the tissue cellular heterogeneity landscape», *Genome Biology*, vol. 18, nº 1, doi :10.1186/s13059-017-1349-1, p. 220, ISSN 1474-760X. URL https://doi.org/10.1186/s13059-017-1349-1. 146

ASHWORTH, A. et collab.. 2011, «Genetic Interactions in Cancer Progression and Treatment», *Cell*, vol. 145, nº 1, doi :10.1016/j.cell.2011.03.020, p. 30–38, ISSN 0092-8674, 1097-4172. URL https://www.cell.com/cell/abstract/S0092-8674(11)00297-2, publisher : Elsevier. 141

AVILA COBOS, F. et collab.. 2018, «Computational deconvolution of transcriptomics data from mixed cell populations», *Bioinformatics*, vol. 34, nº 11, doi :10.1093/bioinformatics/bty019, p. 1969–1979, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/34/11/1969/4813737, publisher : Oxford Academic. 145

BALDUS, S. E. et collab.. 2004, «MUC1 and Nuclear -Catenin Are Coexpressed at the Invasion Front of Colorectal Carcinomas and Are Both Correlated with Tumor Prognosis», *Clinical Cancer Research*, vol. 10, nº 8, doi :10.1158/1078-0432.CCR-03-0163, p. 2790–2796, ISSN 1078-0432, 1557-3265. URL https://clincancerres.aacrjournals.org/content/10/8/2790, publisher : American Association for Cancer Research Section : Regular Articles. 141

BASTIEN, P. et collab.. 2015, «Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data», *Bioinformatics (Oxford, England)*, vol. 31, nº 3, doi :10.1093/bioinformatics/btu660, p. 397–404, ISSN 1367-4811. 148

BIEN, J. et collab.. 2013, «A lasso for hierarchical interactions», *Annals of Statistics*, vol. 41, nº 3, doi :10.1214/13-AOS1096, p. 1111–1141, ISSN 0090-5364, 2168-8966. URL https://projecteuclid.org/euclid.aos/1371150895, publisher : Institute of Mathematical Statistics. 142

CHUN, H. et S. KELEŞ. 2010, «Sparse partial least squares regression for simultaneous dimension reduction and variable selection», *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, vol. 72, nº 1, doi :10.1111/j.1467-9868.2009.00723.x, p. 3–25, ISSN 1369-7412. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810828/. 148

COBOS, F. A., J. ALQUICIRA-HERNANDEZ, J. POWELL, P. MESTDAGH et K. DE PRETER. 2020, «Comprehensive benchmarking of computational deconvolution of transcriptomics data», *bioRxiv*, doi :10.1101/2020.01.10.897116. URL https://www.biorxiv.org/content/early/2020/01/10/2020.01.10.897116, publisher : Cold Spring Harbor Laboratory _eprint : https ://www.biorxiv.org/content/early/2020/01/10/2020.01.10.897116.full.pdf. 145

DE MEULDER, B. et collab.. 2018, «A computational framework for complex disease stratification from multiple large-scale datasets», *BMC systems biology*, vol. 12, nº 1, doi : 10.1186/s12918-018-0556-z, p. 60, ISSN 1752-0509. 148

DERELI, O. et collab.. 2019, «Path2Surv : Pathway/gene set-based survival analysis using multiple kernel learning», *Bioinformatics*, vol. 35, nº 24, doi :10.1093/bioinformatics/btz446, p. 5137–5145, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/35/24/5137/5506626, publisher : Oxford Academic. 142

DOMANY, E. 2014, «Using High-Throughput Transcriptomic Data for Prognosis : A Critical Overview and Perspectives», *Cancer Research*, vol. 74, nº 17, doi :10.1158/0008-5472.CAN-13-3338, p. 4612–4621, ISSN 0008-5472, 1538-7445. URL http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-13-3338. 142, 145

DRIER, Y. et collab.. 2013, «Pathway-based personalized analysis of cancer», *Proceedings of the National Academy of Sciences*, vol. 110, nº 16, doi :10.1073/pnas.1219651110, p. 6388–6393, ISSN 0027-8424, 1091-6490. URL https://www.pnas.org/content/110/16/6388, publisher : National Academy of Sciences Section : Biological Sciences. 145

ELLOUMI, F. et collab.. 2011, «Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples», *BMC medical genomics*, vol. 4, doi :10.1186/1755-8794-4-54, p. 54, ISSN 1755-8794. 145

FA, B. et collab.. 2019, «Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma», *EBioMedicine*, vol. 44, doi : 10.1016/j.ebiom.2019.05.010, p. 250–260, ISSN 23523964. URL https://linkinghub.elsevier.com/retrieve/pii/S2352396419303111. 142, 145

HENDRY, S. et collab.. 2017, «Assessing Tumor-infiltrating Lymphocytes in Solid Tumors : A Practical Review for Pathologists and Proposal for a Standardized Method From the International Immunooncology Biomarkers Working Group : Part 1 : Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma In Situ, Metastatic Tumor Deposits and Areas for Further Research», *Advances in Anatomic Pathology*, vol. 24, nº 5, doi :10.1097/PAP.0000000000000162, p. 235–251, ISSN 1533-4031. 145

HERRMANN, M. et collab.. 2020, «Large-scale benchmark study of survival prediction methods using multi-omics data», *Briefings in Bioinformatics*, doi :10.1093/bib/bbaa167, p. bbaa167, ISSN 1467-5463, 1477-4054. URL https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa167/5895463. 144, 149

HORNUNG, R. et M. N. WRIGHT. 2019, «Block Forests : random forests for blocks of clinical and omics covariate data», *BMC Bioinformatics*, vol. 20, nº 1, doi :10.1186/s12859-019-2942-y, p. 358, ISSN 1471-2105. URL https://doi.org/10.1186/s12859-019-2942-y. 150

HUANG, Z. et collab.. 2019, «SALMON : Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer», *Frontiers in Genetics*, vol. 10, doi :10.3389/fgene.2019.00166, ISSN 1664-8021. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00166/full, publisher : Frontiers. 142, 150

HÄNZELMANN, S. et collab.. 2013, «GSVA : gene set variation analysis for microarray and RNA-Seq data», *BMC Bioinformatics*, vol. 14, nº 1, doi :10.1186/1471-2105-14-7, p. 7, ISSN 1471-2105. URL https://doi.org/10.1186/1471-2105-14-7. 145

KANEHISA, M. et S. GOTO. 2000, «KEGG : kyoto encyclopedia of genes and genomes», *Nucleic Acids Research*, vol. 28, nº 1, doi :10.1093/nar/28.1.27, p. 27–30, ISSN 0305-1048. 142

KANEHISA, M. et collab.. 2019, «New approach for understanding genome variations in KEGG», *Nucleic Acids Research*, vol. 47, nº D1, doi :10.1093/nar/gky962, p. D590–D595, ISSN 1362-4962. 142

LIBERZON, A. et collab.. 2011, «Molecular signatures database (MSigDB) 3.0», *Bioinformatics*, vol. 27, nº 12, doi :10.1093/bioinformatics/btr260, p. 1739–1740, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/27/12/1739/257711, publisher : Oxford Academic. 142

LIM, M. et T. HASTIE. 2015, «Learning interactions via hierarchical group-lasso regularization», *Journal of Computational and Graphical Statistics : A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, vol. 24, nº 3, doi :10.1080/10618600.2014.938812, p. 627–654, ISSN 1061-8600. 142

MAGEN, A. et collab.. 2019, «Beyond Synthetic Lethality : Charting the Landscape of Pairwise Gene Expression States Associated with Survival in Cancer», *Cell Reports*, vol. 28, nº 4, doi :10.1016/j.celrep.2019.06.067, p. 938–948.e6, ISSN 2211-1247. URL http://www.sciencedirect.com/science/article/pii/S2211124719308472. 141

MAO, Y. et collab.. 2018, «Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer», *Cancer Management and Research*, vol. 10, doi :10.2147/CMAR.S171855, p. 3569–3577, ISSN 1179-1322. 146

MILANEZ-ALMEIDA, P. et collab.. 2020, «Cancer prognosis with shallow tumor RNA sequencing», *Nature Medicine*, vol. 26, nº 2, doi :10.1038/s41591-019-0729-3, p. 188–192, ISSN 1078-8956, 1546-170X. URL http://www.nature.com/articles/s41591-019-0729-3. 141

NEWMAN, A. M. et collab.. 2015, «Robust enumeration of cell subsets from tissue expression profiles», *Nature Methods*, vol. 12, nº 5, doi :10.1038/nmeth.3337, p. 453–457, ISSN 1548-7105. URL https://www.nature.com/articles/nmeth.3337, number : 5 Publisher : Nature Publishing Group. 146

PROBST, P. et collab.. 2018, «Hyperparameters and tuning strategies for random forest», *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, doi :10.1002/widm.1301. 148

RAPPOPORT, N. et R. SHAMIR. 2018, «Multi-omic and multi-view clustering algorithms : review and cancer benchmark», *Nucleic Acids Research*, vol. 46, nº 20, doi :10.1093/nar/gky889, p. 10 546–10 562, ISSN 0305-1048. URL https://academic.oup.com/nar/article/46/20/10546/5123392, publisher : Oxford Academic. 149

SHARMA, A. et collab.. 2019, «Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors», *Cell Reports*, vol. 29, nº 8, doi :10.1016/j.celrep.2019.10.045, p. 2164–2174.e5, ISSN 2211-1247. 145

STURM, G. et collab.. 2019, «Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology», *Bioinformatics*, vol. 35, nº 14, doi :10.1093/bioinformatics/btz363, p. i436–i445, ISSN 1367-4803. URL https://academic.oup.com/bioinformatics/article/35/14/i436/5529146, publisher : Oxford Academic. 146

SUBRAMANIAN, A. et collab.. 2005, «Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles», *Proceedings of the National Academy of Sciences*, vol. 102, nº 43, doi :10.1073/pnas.0506580102, p. 15 545–15 550, ISSN 0027-8424, 1091-6490. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0506580102. 142, 145

VASKE, C. J. et collab.. 2010, «Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM», *Bioinformatics*, vol. 26, nº 12, doi :10.1093/bioinformatics/btq182, p. i237–i245, ISSN 1367-4803. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2881367/. 145

DE VLAMING, R. et P. J. F. GROENEN. 2015, «The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics», *BioMed Research International*, doi :https://doi.org/10.1155/2015/143712. URL https://www.hindawi.com/journals/bmri/2015/143712/, iSSN : 2314-6133 Pages : e143712 Publisher : Hindawi Volume : 2015. 142

VOLKMANN, A., R. DE BIN, W. SAUERBREI et A.-L. BOULESTEIX. 2019, «A plea for taking all available clinical information into account when assessing the predictive value of omics data», *BMC Medical Research Methodology*, vol. 19, nº 1, doi :10.1186/s12874-019-0802-0, p. 162, ISSN 1471-2288. URL https://doi.org/10.1186/s12874-019-0802-0. 143

WANG, L. et collab.. 2014, «A Modified Adaptive Lasso for Identifying Interactions in the Cox Model with the Heredity Constraint», *Statistics & probability letters*, vol. 93, doi : 10.1016/j.spl.2014.06.024, p. 126–133, ISSN 0167-7152. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111275/. 142

WANG, Y.-P. et Q.-Y. LEI. 2018, «Metabolic recoding of epigenetics in cancer», *Cancer Communications (London, England)*, vol. 38, nº 1, doi :10.1186/s40880-018-0302-3, p. 25, ISSN 2523-3548. 149

WU, C. et collab.. 2019, «A Selective Review of Multi-Level Omics Data Integration Using Variable Selection», *High-Throughput*, vol. 8, nº 1, doi :10.3390/ht8010004, p. 4. URL https://www.mdpi.com/2571-5135/8/1/4, number : 1 Publisher : Multidisciplinary Digital Publishing Institute. 149

WU, M. et collab.. 2018, «Identifying gene-gene interactions using penalized tensor regression», *Statistics in Medicine*, vol. 37, nº 4, doi :10.1002/sim.7523, p. 598–610, ISSN 1097-0258. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7523, _eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7523. 141, 142

ZHANG, N. et collab.. 2016, «Classification of cancers based on copy number variation landscapes», *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1860, nº 11, Part B, doi :10.1016/j.bbagen.2016.06.003, p. 2750–2755, ISSN 0304-4165. URL http://www.sciencedirect.com/science/article/pii/S0304416516302082. 149

ZHAO, Q. et collab.. 2015, «Combining multidimensional genomic measurements for predicting cancer prognosis : observations from TCGA», *Briefings in Bioinformatics*, vol. 16, nº 2, doi :10.1093/bib/bbu003, p. 291–303, ISSN 1467-5463. URL https://academic.oup.com/bib/article/16/2/291/246070, publisher : Oxford Academic. 149

ZHENG, X. et collab.. 2020, «Comparison of pathway and gene-level models for cancer prognosis prediction», *BMC Bioinformatics*, vol. 21, nº 1, doi :10.1186/

s12859-020-3423-z, p. 76, ISSN 1471-2105. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-3423-z. 142, 145

ZHU, B. et collab.. 2017, «Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers», *Scientific Reports*, vol. 7, nº 1, doi :10.1038/s41598-017-17031-8, p. 16 954, ISSN 2045-2322. URL https://www.nature.com/articles/s41598-017-17031-8, number : 1 Publisher : Nature Publishing Group. 150

Annexe A

Figures et Tableaux Annexes

A.1 Chapitre 2

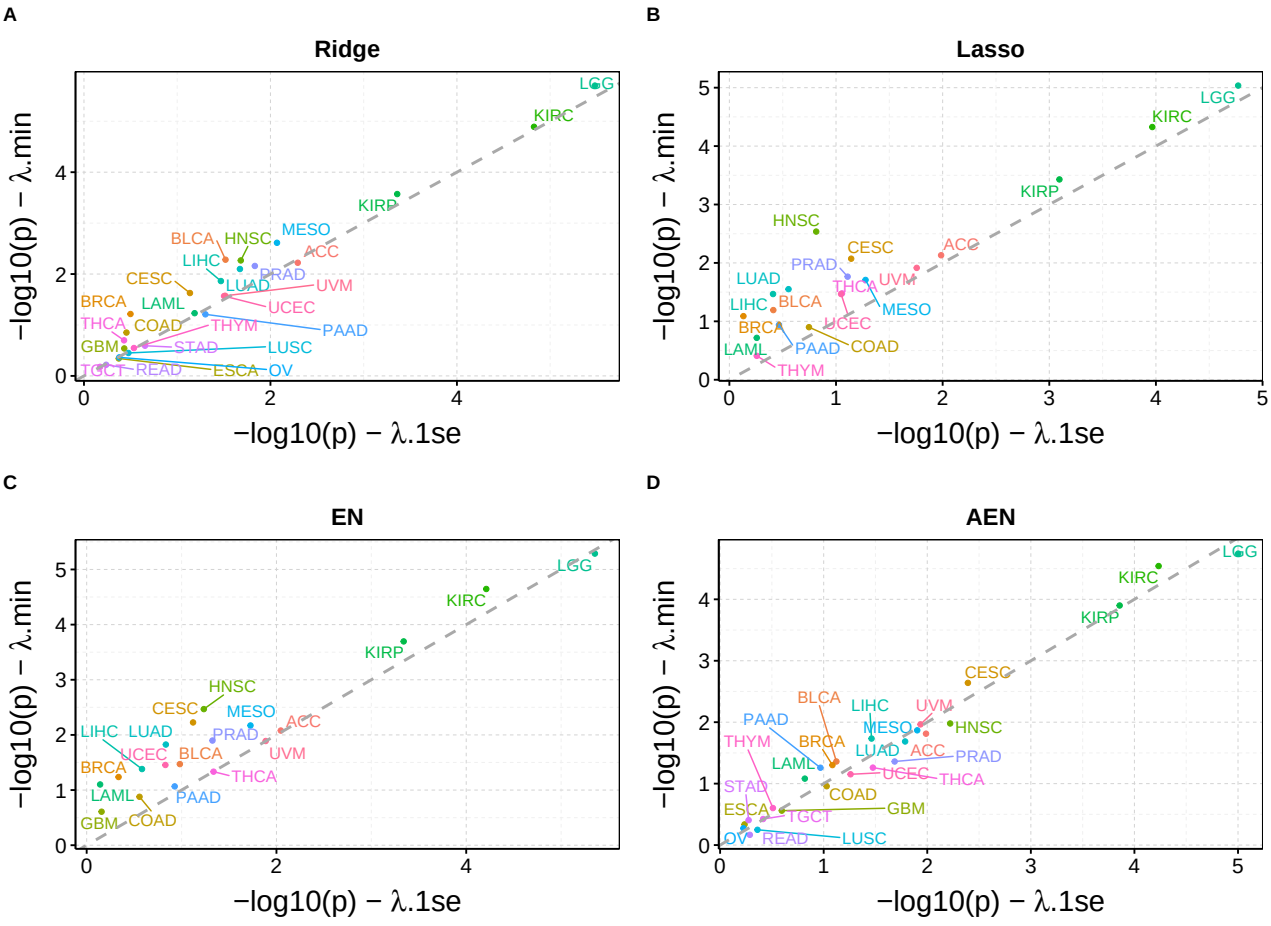


FIGURE A.1 – **P-valeurs médianes du modèle de Cox univarié ($-log_{10}$) obtenues avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés.**
Nous avons calculé les indices pronostiques par 10 répétitions d'une validation croisée (K=5). La médiane des 50 p-valeurs du modèle de Cox univarié ($-log_{10}$) est affichée ici pour chaque cancer.
(A) Ridge, (B) lasso, (C) elastic net (EN), (D) adaptive elastic net (AEN).

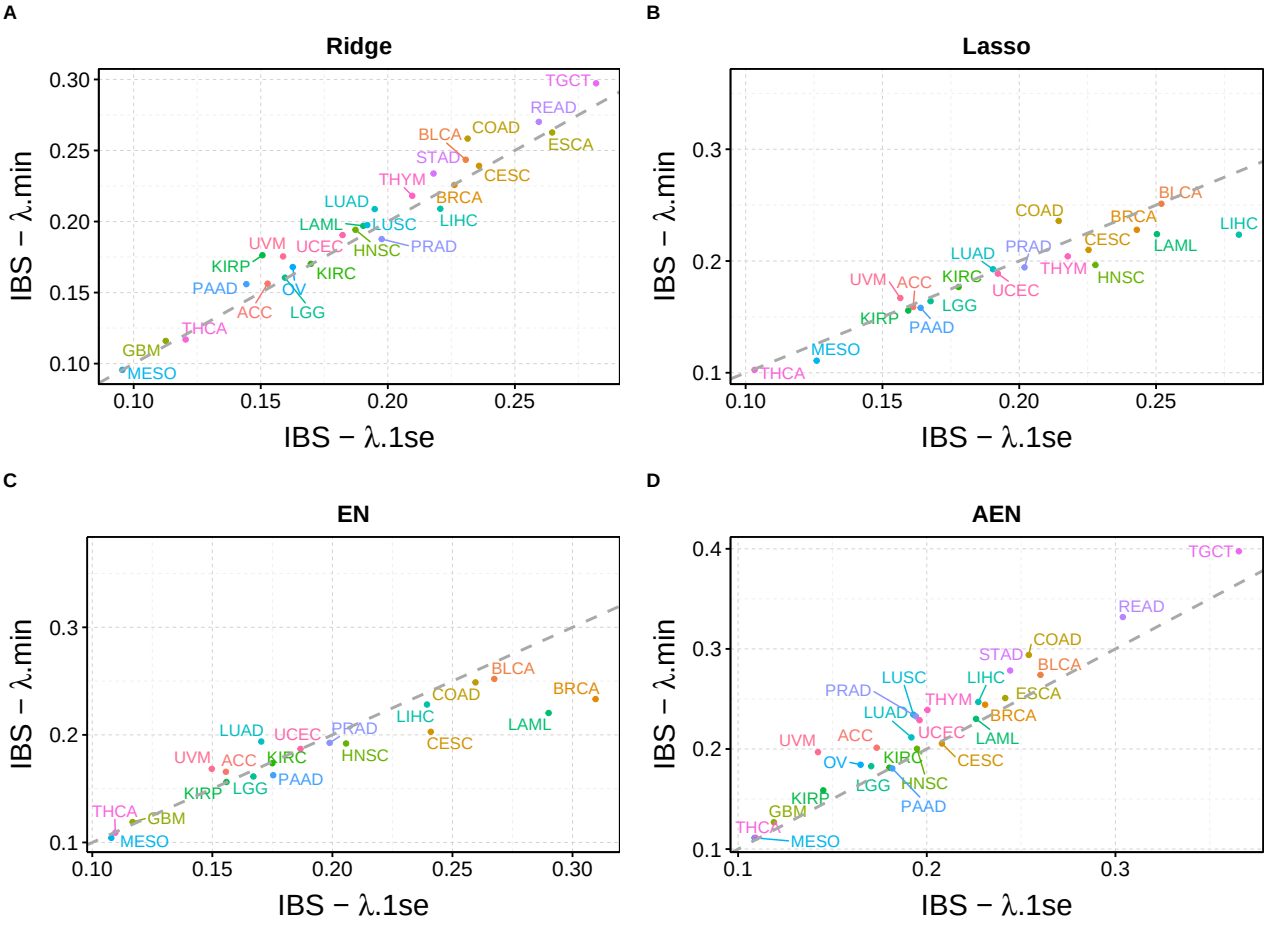


FIGURE A.2 – **IBS obtenus avec λ.min et λ.1se pour les quatre pénalisations et l'ensemble des 26 cancers de TCGA étudiés.**

Nous avons calculé les indices pronostiques par 10 répétitions d'une validation croisée (K=5). La médiane des 50 IBS est affichée ici pour chaque cancer.
(A) Ridge, (B) lasso, (C) elastic net (EN), (D) adaptive elastic net (AEN).

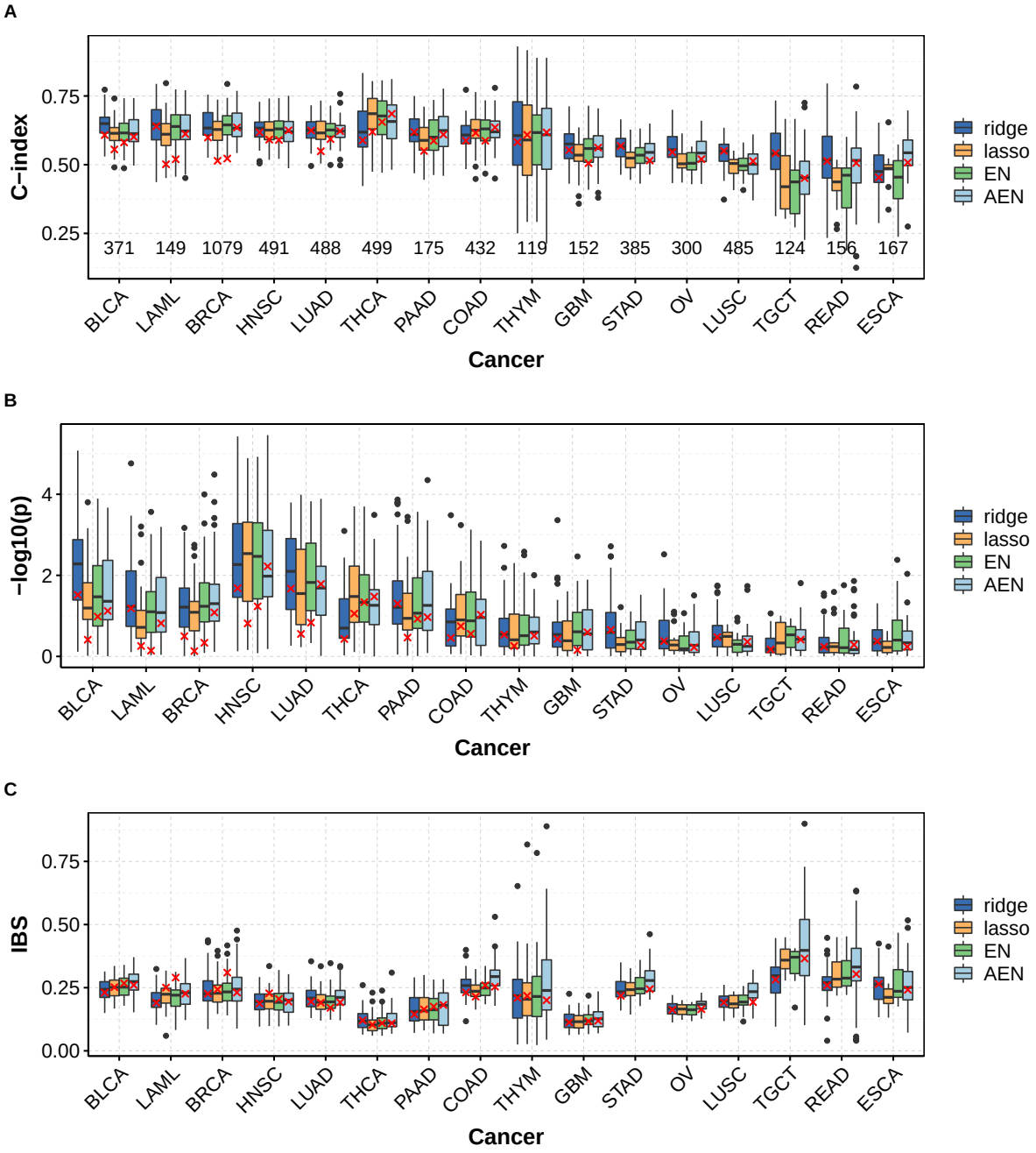


FIGURE A.3 – **C-index (A), p-valeurs (B), et IBS (C) obtenus avec les données mRNA-seq.**
Pour chaque cancer et chaque méthode de pénalisation, nous calculons les 50 C-index, p-valeurs, et IBS par 10 répétitions d'une validation croisée (K=5) sur l'ensemble des gènes. Les boites représentent les C-index (A), p-valeurs (-log10) (B), et IBS (C) obtenus avec λ_{min}, et les croix rouges représentent le C-index (A), p-valeurs (-log10) (B), et IBS (C) médians obtenus avec λ_{1se}.
Les nombres de patients dans les jeux de données sont notés en noir en bas du graphique A.
Les 16 cancers qui ont le plus faible C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 10 autres cancers étudiés sont dans le corps du texte (Fig. 2.3).

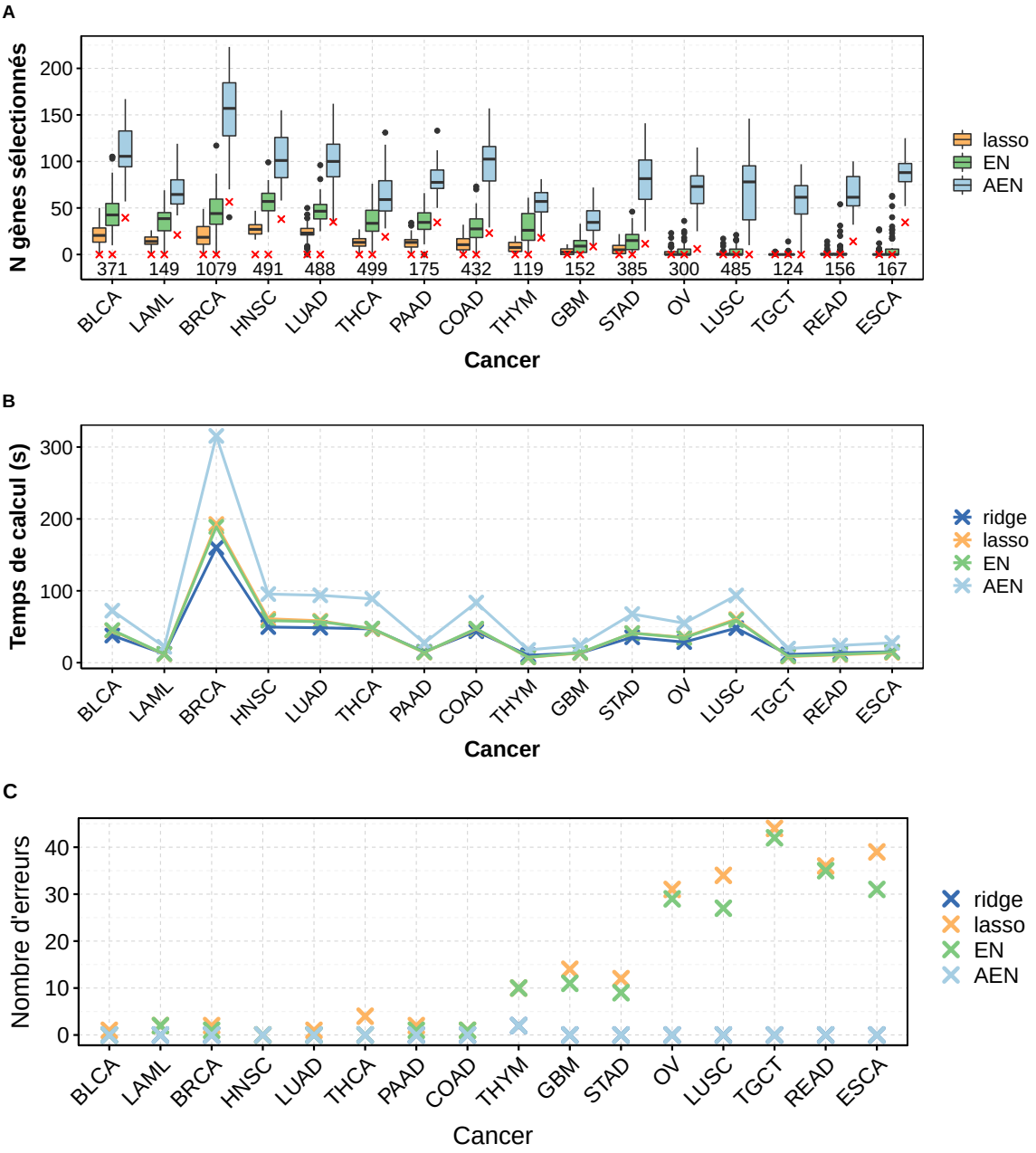


FIGURE A.4 – **Nombre de gènes sélectionnés (A), temps de calcul médian (s) (B), et nombre d'erreurs dans le calcul du C-index avec λ_{min} (C) des méthodes de pénalisation obtenus avec les données mRNA-seq.**

Pour chaque cancer et chaque méthode de pénalisation, nous apprenons 50 modèles par 10 répétitions d'une validation croisée (K=5) sur l'ensemble des gènes. Pour chacun des 50 apprentissages, le nombre de gènes sélectionnés (A) et le temps de calcul (B) sont calculés. Les croix rouges dans le graphique A représentent le nombre de gènes sélectionnés avec λ_{1se} médian.

Les nombres de patients dans les jeux de données sont notés en noir en bas du graphique A.

Les 16 cancers qui ont le plus faible C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 10 autres cancers étudiés sont dans le corps du texte (Fig. 2.4).

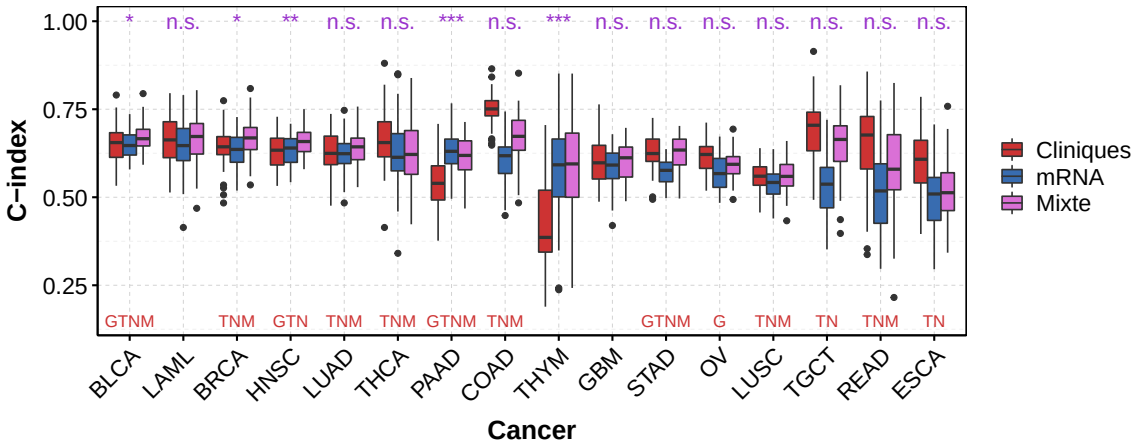


FIGURE A.5 – **C-index obtenus avec les données cliniques, mRNA-seq, et en mixant les deux types de données.**
Pour chaque cancer, 50 C-index sont calculés en utilisant 6 variables cliniques classiques (âge, genre, grade, T, N, M - rouge) lorsqu'elles sont disponibles, les données mRNA-seq (bleu), et en mixant les deux types de données (*i.e.* les 6 variables cliniques et l'indice pronostique calculé avec les données mRNA-seq sont utilisés - violet). Pour chaque cas, les 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5).
L'âge est disponible pour tous les cancers, le genre pour les 8 cancers non unisexe (*i.e.* CESC et UCEC sont des cancers touchant uniquement les femmes, et PRAD uniquement les hommes). Les lettres rouges en bas du graphique indique la présence ou non des autres variables cliniques (G : grade, T : « *Tumor* », N : « *Node* », M : « *Metastasis* »).
Les p-valeurs d'un test de Wilcoxon unilatéral corrigées par la méthode de Benjamini-Hochberg suivant l'ensemble des 26 cancers permettant d'observer si le C-index médian obtenu avec le mixte des données cliniques et miRNA-seq (violet) est significativement supérieur à celui obtenu avec uniquement les données cliniques (rouge) sont indiquées en violet sous forme d'étoiles suivant le niveau de significativité (légende ci-dessous).
n.s : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

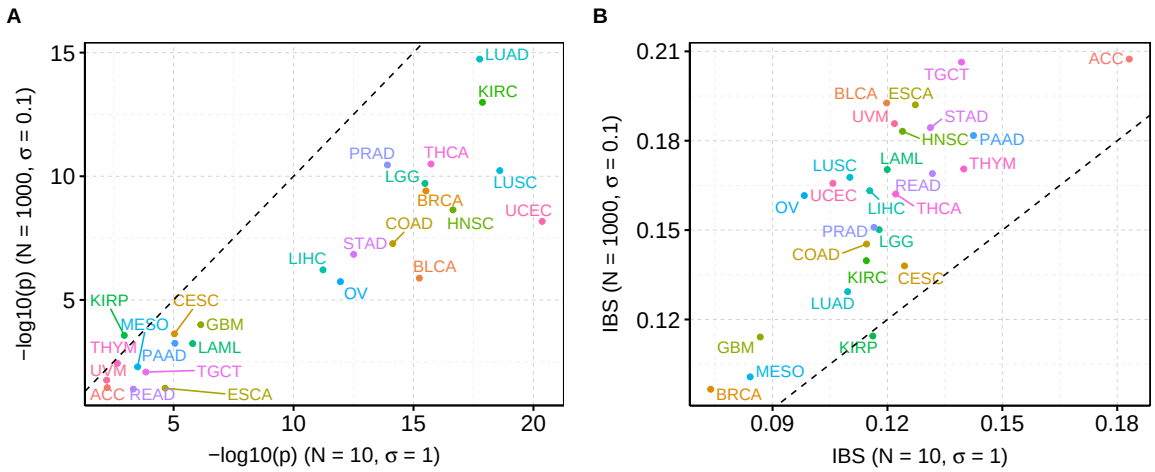


FIGURE A.6 – **Médianes des p-valeurs du modèle de Cox univarié (-log10) (A) et des IBS (B) obtenus pour différents paramètres de simulations contenant la même quantité d'information pour les 26 cancers de TCGA étudiés - données simulées.**
Les C-index médians sont obtenus avec N = 1 000 et σ = 0.1 en ordonnée, et avec N = 10 et σ = 1 en abscisse. La droite d'équation $y = x$ est tracée en pointillée.

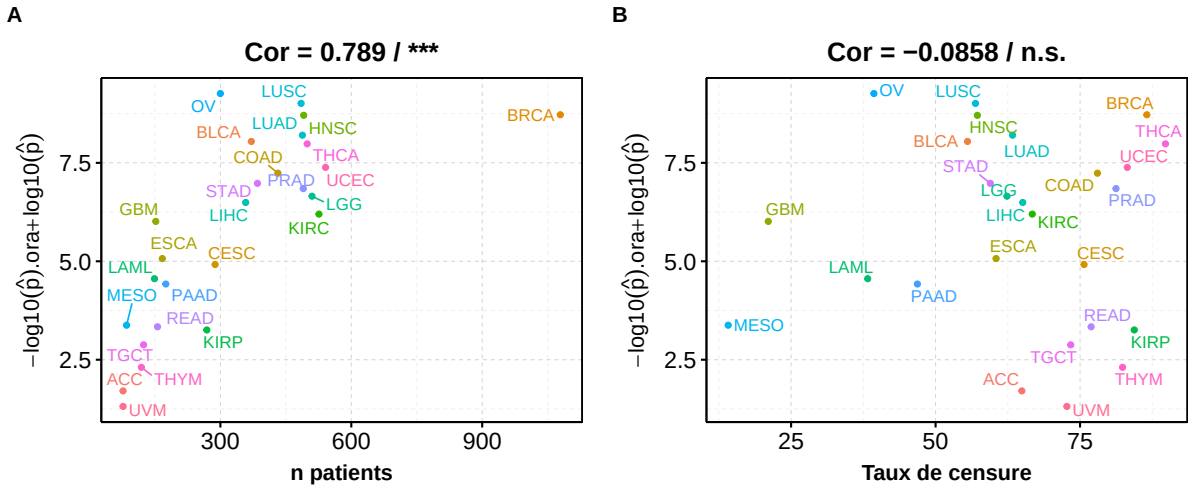


FIGURE A.7 – **Différences médianes entre les p-valeurs oracles (-log10) et estimées en fonction du nombre de patients (A) et du taux de censure (B) pour l'ensemble des 26 cancers étudiés.**
Les corrélations et les niveau de significativité d'un test de Spearman sont indiqués en haut des graphiques. (*** : p-valeur < 0,001 ; n.s. : non significatif).
Les différences médianes sont calculées sur l'ensemble des simulations (σ = 0,1 ; 0,25 ; 0,5 ; 1 et N = 10 ; 25 ; 50 ; 100 ; 1 000).

TABLEAU A.1 – **Différences médianes entre les p-valeurs oracles du modèle de Cox univarié (-log10) et estimés par le modèle de Cox pénalisé pour les 26 cancers de TCGA étudiés.**
Les médianes sont calculées pour chaque paramètres de simulation ($\sigma = 0,1;0,25;0,5;1$ et N = 10;25;50;100).

Cancer	ridge	lasso	EN	AEN
ACC	1.4	1.8	1.7	1.5
BLCA	9.9	7.2	8	7.1
BRCA	11	7.5	8.7	7.5
CESC	4.8	5.2	4.9	4.3
COAD	8	6.6	7.2	6.1
ESCA	4.9	5.4	5.1	4.3
GBM	6	5.8	6	5.4
HNSC	12	7.7	8.7	7.2
KIRC	7.9	5.8	6.2	5.5
KIRP	3	3.4	3.3	2.9
LAML	4.5	4.6	4.6	4.3
LGG	8.9	6	6.7	5.9
LIHC	7.2	6.2	6.5	5.7
LUAD	10	7.4	8.2	7
LUSC	12	7.8	9	8
MESO	3.2	3.5	3.4	3.2
OV	11	8.3	9.3	7.8
PAAD	4.2	4.5	4.4	3.8
PRAD	7.8	6.4	6.8	5.9
READ	3	3.4	3.3	2.9
STAD	8.3	6.2	7	6.1
TGCT	2.5	3	2.9	2.5
THCA	9	7.1	8	6.4
THYM	1.9	2.4	2.3	1.9
UCEC	9.1	6.5	7.4	6.1
UVM	1.1	1.4	1.3	1.2

TABLEAU A.2 – **Différences moyennes entre les IBS estimés par le modèle de Cox pénalisé et les IBS orocles pour les 16 cancers de TCGA étudiés.**
Les médianes sont calculées pour chaque paramètres de simulation ($\sigma = 0,1;0,25;0,5;1$ et $N = 10;25;50;100$).

Cancer	ridge	lasso	EN	AEN
ACC	0.067	0.089	0.083	0.099
BLCA	0.067	0.05	0.053	0.054
BRCA	0.036	0.021	0.025	0.027
CESC	0.062	0.065	0.062	0.067
COAD	0.06	0.047	0.05	0.055
ESCA	0.082	0.085	0.082	0.085
GBM	0.046	0.043	0.044	0.043
HNSC	0.064	0.04	0.046	0.046
KIRC	0.045	0.033	0.034	0.036
KIRP	0.046	0.047	0.047	0.051
LAML	0.058	0.06	0.058	0.061
LGG	0.046	0.033	0.037	0.038
LIHC	0.06	0.053	0.054	0.057
LUAD	0.061	0.041	0.045	0.047
LUSC	0.071	0.044	0.048	0.052
MESO	0.032	0.035	0.033	0.037
OV	0.067	0.046	0.051	0.048
PAAD	0.06	0.062	0.06	0.061
PRAD	0.048	0.038	0.04	0.043
READ	0.076	0.081	0.081	0.091
STAD	0.063	0.046	0.049	0.054
TGCT	0.07	0.073	0.073	0.083
THCA	0.051	0.037	0.041	0.038
THYM	0.053	0.063	0.058	0.065
UCEC	0.057	0.035	0.041	0.044
UVM	0.057	0.074	0.07	0.082

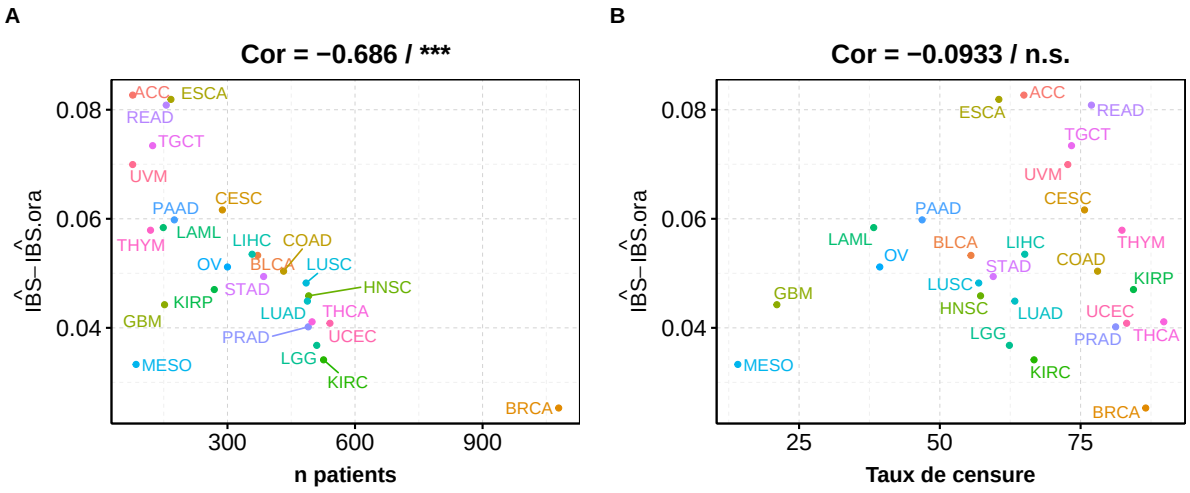


FIGURE A.8 – **Différences médianes entre les IBS estimés et oracles en fonction du nombre de patients (A) et du taux de censure (B) pour l'ensemble des 26 cancers étudiés.**
Les corrélations et les niveau de significativité d'un test de Spearman sont indiqués en haut des graphiques. (*** : p-valeur < 0,001 ; n.s. : non significatif).
Les différences médianes sont calculées sur l'ensemble des simulations ($\sigma = 0,1; 0,25; 0,5; 1$ et $N = 10; 25; 50; 100; 1\,000$).

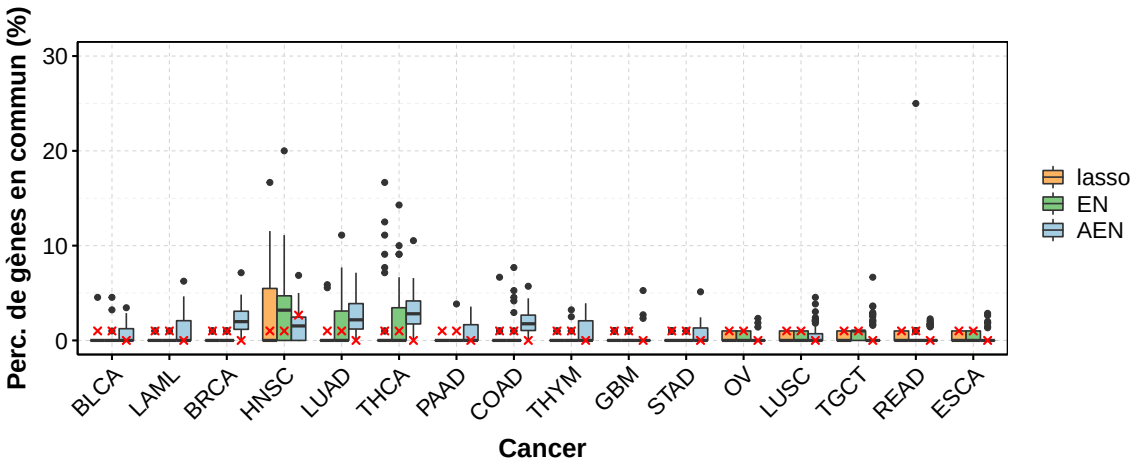


FIGURE A.9 – **Stabilité des gènes sélectionnés.**
Pourcentage de gènes sélectionnés en commun dans deux sous-jeux de données comportant chacun 50% des patients. Les *boxplots* sont les résultats obtenus avec $\lambda_{\min}$, et les croix rouges sont les résultats obtenus avec λ_{1se}.
Les 16 cancers qui ont le plus faible C-index médian calculé avec la pénalisation ridge sont présentés. Les résultats obtenus pour les 10 autres cancers étudiés sont placés dans le corps du texte (Fig. 2.14).

TABLEAU A.3 – **Différences médianes entre les p-valeurs (-log10) obtenues si la vérité terrain était connue et les p-valeurs (-log10) estimées pour les 26 cancers de TCGA et les quatre pénalisations étudiées.**
EN : elastic net ; AEN : adaptive elastic net.

Cancer	ridge	lasso	EN	AEN
ACC	0.17	0.28	0.33	0.22
BLCA	4.9	2.8	3.6	2.9
BRCA	6.1	3	3.9	3.1
CESC	1.7	1.6	1.6	1.4
COAD	3.1	2.2	2.6	2
ESCA	1.7	1.6	1.9	1.2
GBM	1.6	1.7	2.1	1.5
HNSC	6.6	3.4	4.3	3.4
KIRC	3.7	2.2	2.9	2.5
KIRP	0.95	1.1	1.2	0.95
LAML	0.95	1.2	1.4	0.91
LGG	4	2.5	3	2.5
LIHC	3	2	2.5	2
LUAD	5.3	3.2	3.9	3.2
LUSC	7.6	3.5	4.7	3.6
MESO	0.57	0.74	0.79	0.67
OV	6	3.3	4.5	3.2
PAAD	1.4	1	1.2	1.1
PRAD	2.8	2.1	2.3	2
READ	0.83	0.84	0.89	0.7
STAD	4.4	2.6	3.5	2.8
TGCT	0.31	0.61	0.71	0.53
THCA	4.6	3.2	4	2.9
THYM	0.19	0.33	0.3	0.26
UCEC	3.2	2.4	2.7	2.2
UVM	0.064	0.2	0.2	0.18

TABLEAU A.4 – **Différences médianes entre les IBS obtenus si la vérité terrain était connue et les IBS estimées pour les 26 cancers de TCGA et les quatre pénalisations étudiées.**
EN : elastic net ; AEN : adaptive elastic net.

Cancer	ridge	lasso	EN	AEN
ACC	0.017	0.025	0.029	0.041
BLCA	0.043	0.024	0.03	0.033
BRCA	0.025	0.01	0.014	0.016
CESC	0.033	0.024	0.027	0.033
COAD	0.032	0.019	0.024	0.027
ESCA	0.039	0.033	0.036	0.037
GBM	0.017	0.015	0.017	0.018
HNSC	0.042	0.019	0.025	0.025
KIRC	0.028	0.015	0.019	0.021
KIRP	0.018	0.019	0.019	0.024
LAML	0.021	0.019	0.023	0.022
LGG	0.025	0.015	0.019	0.02
LIHC	0.035	0.021	0.026	0.03
LUAD	0.038	0.019	0.025	0.026
LUSC	0.051	0.021	0.028	0.029
MESO	0.0085	0.0093	0.0093	0.013
OV	0.041	0.02	0.025	0.025
PAAD	0.021	0.016	0.02	0.024
PRAD	0.023	0.015	0.018	0.022
READ	0.03	0.028	0.031	0.041
STAD	0.04	0.022	0.027	0.032
TGCT	0.018	0.025	0.029	0.039
THCA	0.028	0.016	0.022	0.02
THYM	0.0098	0.014	0.013	0.023
UCEC	0.031	0.017	0.021	0.024
UVM	0.0078	0.019	0.017	0.026

A.2 Chapitre 3

TABLEAU A.5 – **Augmetation médiane de l'opposé du log10 de la p-valeur du modèle de Cox univarié après pré-filtrage et niveau de significativité pour les 26 cancers étudiés.**
Les p-valeurs sont calculées par un test des rangs signés de Wilcoxon unilatéral, et corrigées par la méthode de Benjamini-Hochberg pour chacune des méthodes.

cancer	ridge	lasso	EN	AEN	cancer	ridge	lasso	EN	AEN
ACC	0.29	0	0.11	0.29	LGG	0.14	0.41	0.67	1
	***	n.s.	n.s.	**		n.s.	**	***	***
BLCA	0.074	0.52	0.35	0.21	LIHC	0.096	0.15	0.15	0.17
	n.s.	***	***	+		*	*	*	n.s.
BRCA	0.41	0.62	0.57	0.52	LUAD	0.01	0.29	0	0.15
	***	***	***	***		n.s.	*	n.s.	n.s.
CESC	0.12	0.65	0.034	0	MESO	0	0.29	0.14	0.37
	*	***	n.s.	n.s.		n.s.	**	*	**
HNSC	0.2	0.064	0.27	0.22	PAAD	0.23	0.53	0.25	0.14
	n.s.	n.s.	*	n.s.		***	***	*	n.s.
KIRC	0.17	0.31	0.12	0.48	PRAD	0.044	0.28	0	0.46
	+	+	*	**		n.s.	**	n.s.	***
KIRP	0	0	0.063	0.039	UCEC	0.13	0.065	0.09	0.13
	n.s.	n.s.	*	+		n.s.	n.s.	n.s.	n.s.
LAML	0.17	0.5	0.41	0.31	UVM	0.22	0	0.058	0
	**	***	***	**		**	n.s.	+	n.s.

TABLEAU A.6 – **Diminution médiane de l'IBS après pré-filtrage et niveau de significativité pour les 26 cancers étudiés.**
Les p-valeurs sont calculées par un test des rangs signés de Wilcoxon unilatéral, et corrigées par la méthode de Benjamini-Hochberg pour chacune des méthodes.

cancer	ridge	lasso	EN	AEN	cancer	ridge	lasso	EN	AEN
ACC	0.017	0.0038	0.0023	0.05	LGG	0	0.0037	0.00049	0.015
	***	n.s.	n.s.	***		n.s.	n.s.	n.s.	***
BLCA	0.011	0.022	0.017	0.045	LIHC	0.0075	0	0.012	0.027
	*	***	***	***		n.s.	n.s.	*	***
BRCA	0.017	0.014	0.015	0.035	LUAD	0.015	0	0	0.029
	**	*	**	***		***	n.s.	n.s.	***
CESC	0.0099	0	0	0	MESO	0	0.0029	0	0.0042
	**	n.s.	n.s.	n.s.		n.s.	n.s.	n.s.	+
HNSC	0.0084	0.013	0.00057	0.013	PAAD	0.011	0.013	0.013	0.023
	*	*	n.s.	**		**	*	**	***
KIRC	0	0.0053	0.0037	0.0077	PRAD	0.0056	0.0061	0.0024	0.031
	n.s.	+	***	*		**	n.s.	n.s.	***
KIRP	0.0037	0	0.0022	0	UCEC	0.019	0.011	0.034	0.045
	+	n.s.	n.s.	n.s.		**	+	***	***
LAML	0.012	0.029	0.033	0.04	UVM	0.025	0.012	0.019	0.041
	*	***	***	***		***	n.s.	**	**

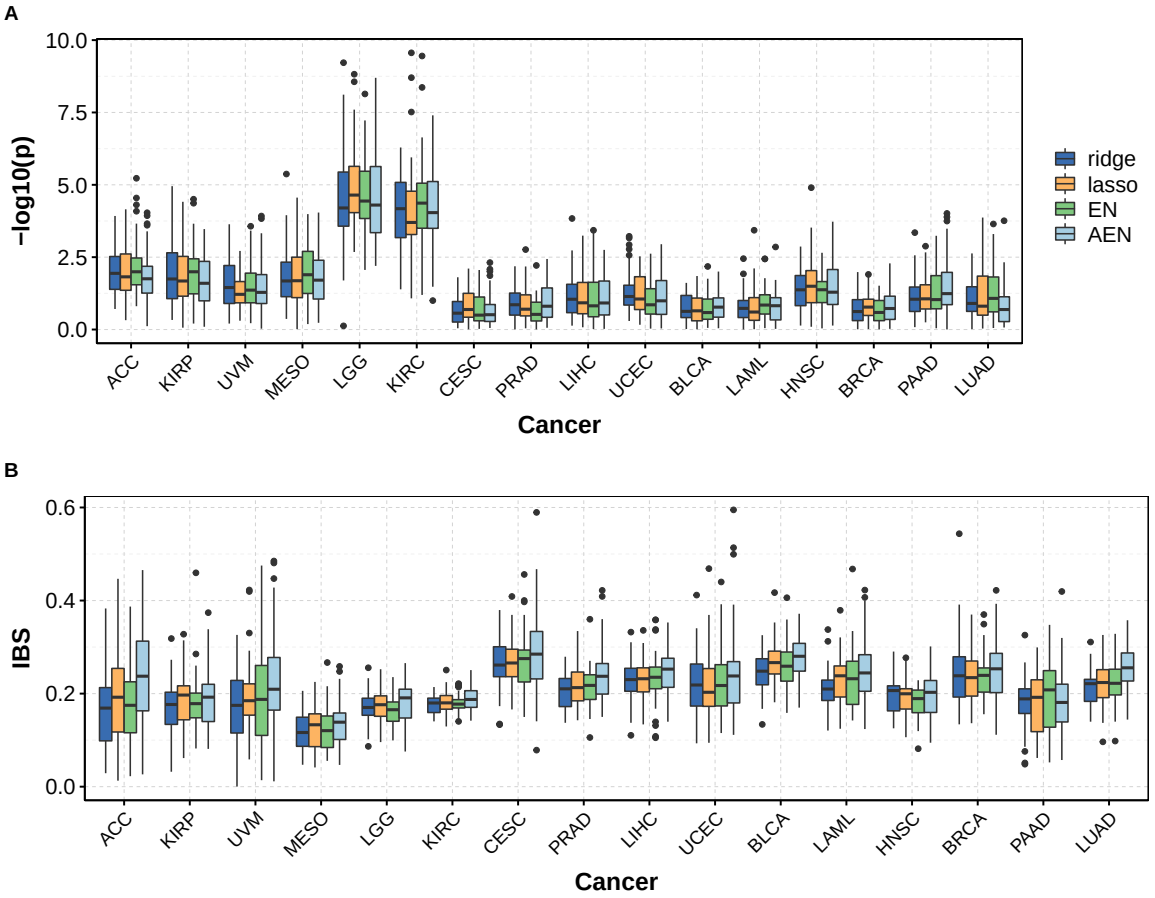


FIGURE A.10 – **P-valeurs (-log10) (A) et IBS (B) obtenus après prè-filtrage pour les quatre méthodes de pénalisation (*i.e.* ridge, lasso, elastic net et adaptive elastic net) et 16 cancers de TCGA.** Les cancers sont classés par ordre décroissant du C-index médian calculé avec la pénalisation ridge.

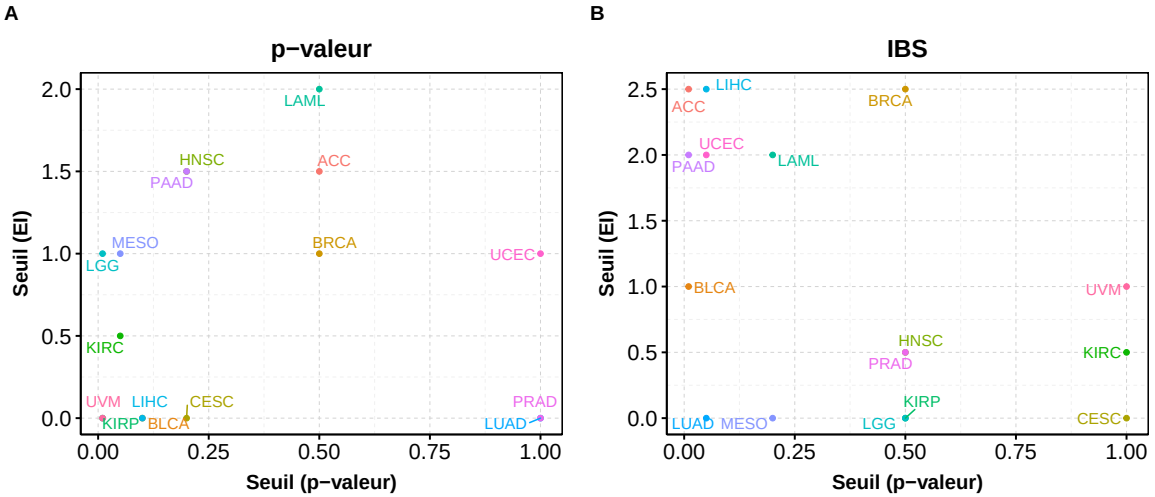


FIGURE A.11 – **Seuils optimaux pour la pénalisation elastic net, l'ensemble des 26 cancers, et la p-valeur du modèle de Cox univarié (A) et l'IBS (B).**

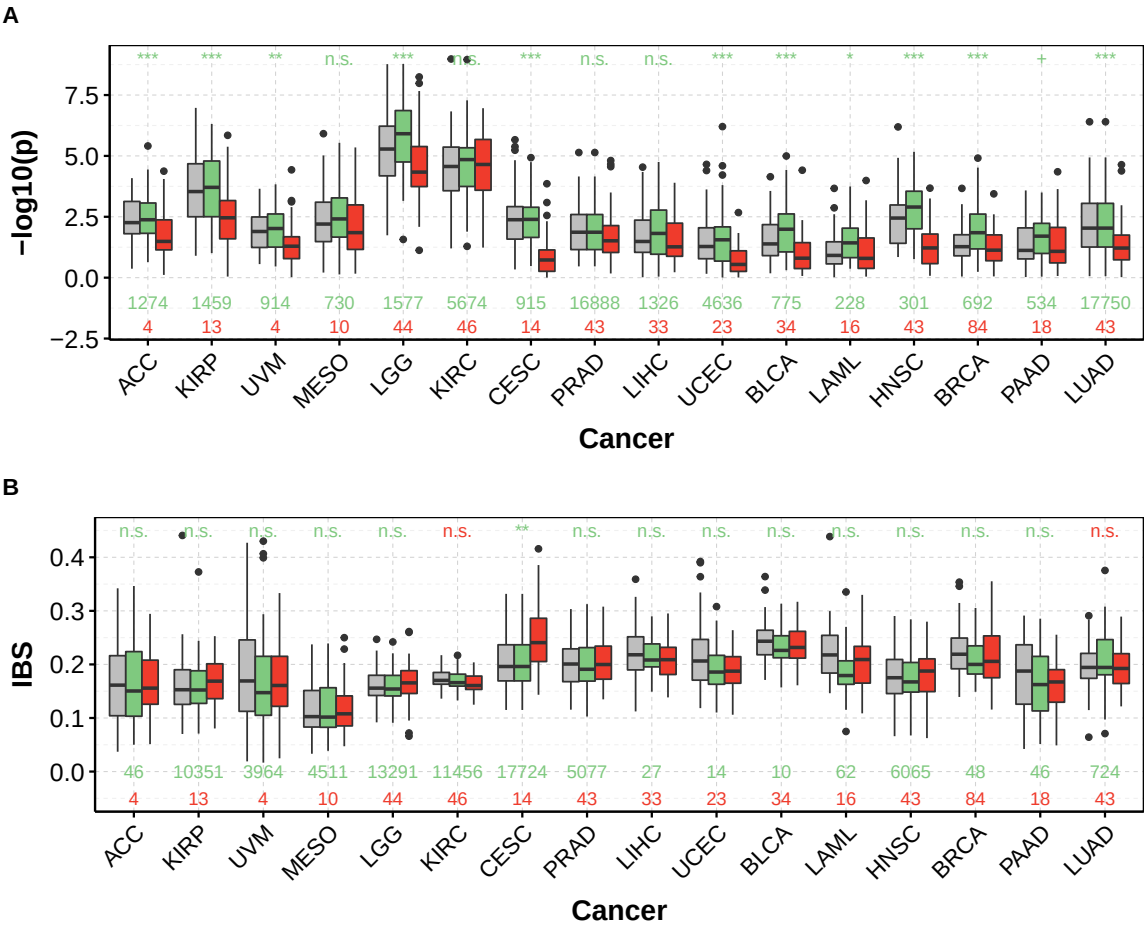


FIGURE A.12 – **P-valeurs du modèle de Cox univarié (-log10) (A) et IBS (B) obtenus sans pré-filtrage (gris), avec le pré-filtrage bi-dimensionnel et la pénalisation elastic net (vert), et avec l'algorithme ISIS (rouge) pour 16 cancers de TCGA.**

Dans chaque cas, les métriques sont calculées par 10 répétitions d'une validation croisée (K=5). Les nombres rouges et verts indiquent le nombre de gènes après filtrage par la méthode ISIS et bi-dimensionnelle, respectivement. Les p-valeurs corrigées par la méthodes de Benjamini-Hochberg permettant de tester si la métrique médiane obtenue avec les deux méthodes sont significativement différente (vert : les prédictions médianes sont meilleures avec le filtrage bi-dimensionnel, rouge : les prédictions médianes sont meilleures avec la méthode ISIS).

n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

A.3 Chapitre 4

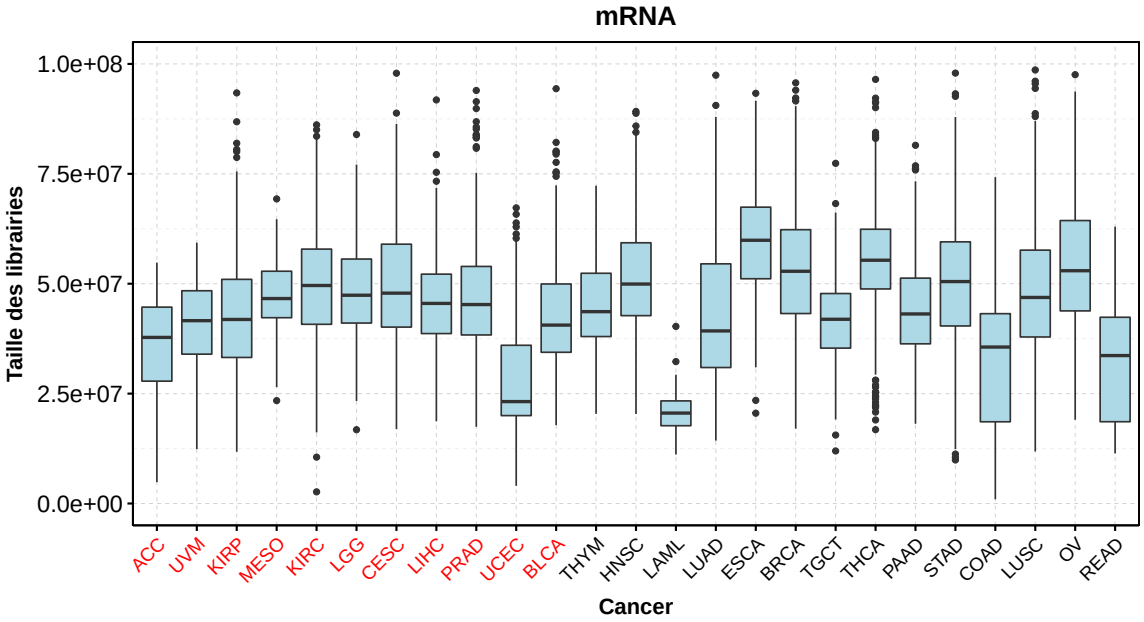


FIGURE A.13 – **Tailles des librairies pour 25 cancers de TCGA obtenues avec les données RNA-seq pour les mARN.**
Les 11 cancers dont le nom est écrit en rouge sont ceux qui sont étudiés dans le chapitre 4.

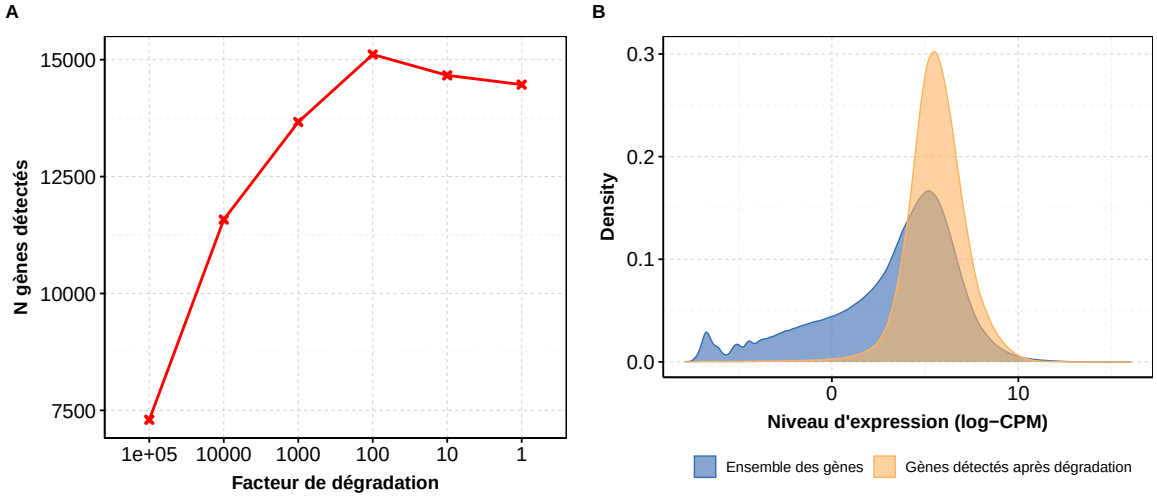


FIGURE A.14 – **Impact de la dégradation de la taille des banques des données mRNA-seq sur le nombre de gènes détectés (A) et leur niveau d'expression (B) pour KIRC.**
(A) Les données brutes sont dégradées suivant 5 facteurs différents (1, 10, 100, 1 000, 10 000, 100 000 - abscisse), et le nombre de gènes détectés est calculé (ordonnée).
(B) Densité (estimation de la distribution) du niveau d'expression de l'ensemble des gènes (bleu), et des gènes détectes après dégradation d'un facteur 10 000 (jaune) pour l'ensemble des patients. Les données de comptage sont normalisées (log2-CPM) dans les deux cas.

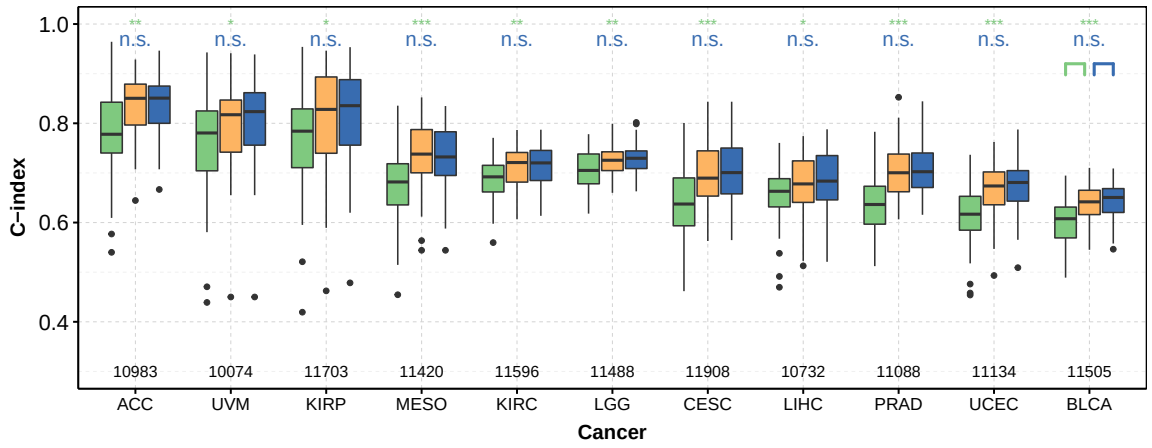


FIGURE A.15 – **Cause de la diminution des capacités de prédiction induite par la dégradation des données mRNA-seq des 11 cancers étudiés.**

Boxplot des C-index obtenus avec les données miRNA-seq pour un facteur de dégradation de 10 000 (vert), sans dégradation (bleu), et sans dégradation mais avec uniquement les gènes détectés après dégradation d'un facteur 10 000 (jaune). Le nombre de gènes détectés après dégradation d'un facteur 10 000 est indiqué en noir en bas du graphique pour chaque cancer. Pour chaque scénario, 50 C-index sont calculés par 10 répétitions d'une validation croisée (K=5).

Les p-valeurs corrigées par la méthode de Benjamini-Hochberg d'un test de Wilcoxon unilatéral permettant de tester si le C-index médian obtenu dans le « scénario jaune » est significativement plus grand que celui obtenu dans le « scénario vert » sont indiquées sous forme d'étoiles vertes en haut du graphique.

Les p-valeurs corrigées par la méthode de Benjamini-Hochberg d'un test de Wilcoxon unilatéral permettant de tester si le C-index médian obtenu dans le « scénario bleu » est significativement plus grand que celui obtenu dans le « scénario jaune » sont indiquées sous forme d'étoiles bleues en haut du graphique.

n.s. : non-significatif ; + : p-valeur < 0,1 ; * : p-valeur < 0,05 ; ** : p-valeur < 0,01 ; *** : p-valeur < 0,001.

Annexe B

Liste des acronymes

ADN acide désoxyribonucléique. 6, I

AEN Adaptive Elastic Net. 29

ARN acide ribonucléique. 6, I

BS score de Brier - *Brier Score* -. 25

CPM *Count Per Million*. 12

EI Écart Interquartile. 93

EN Elastic Net. 29

FDP *False Discovery Proportion*. 48

IBS score de Brier intégré - *Integrated Brier Score* -. 26

INCa Institut National du Cancer. 3, 4, XIX

INSERM Institut National de la Santé et de la Recherche Médicale. 5, XIX

ISIS *Iterative Sure Independance Screening*. 102

NCI *National Cancer Institute*. 15

NHGRI *National Human Genome Research Institute*. 15

OMS Organisation Mondiale de la Santé. 3

PI indice pronostique - *prognostic index* -. 21, 23

TCGA *The Cancer Genome Atlas*. 15

VST *Variance Stabilizing Transformation*. 95

Annexe C

Glossaire

biomarqueur (définition INSERM) Molécule (enzyme, hormone, métabolite, etc.), voire un type de cellule, dont la présence ou la concentration anormale dans le sang ou les urines signale un évènement ou un statut physiologique particulier. 5

cancer (définition (INCa) Maladie provoquée par la transformation de cellules qui deviennent anormales et prolifèrent de façon excessive. Ces cellules déréglées finissent par former une masse qu'on appelle tumeur maligne. 4

censure donnée de survie pour laquelle l'évènement étudié (i.e., décès du patient ou nouvel évènement associé à la tumeur) n'est pas observée au cours du suivi. 17, I

hétérogénéité inter-tumorale hétérogénéité des caractéristiques génétiques et phénotypiques des tumeurs entre les patients. 4

hétérogénéité intra-tumorale hétérogénéité des caractéristiques génétiques et phénotypiques au sein d'une même tumeur. 4

malédiction de la grande dimension apparition de phénomènes qui ont lieu lorsque le nombre p de variables excède le nombre n de patients. La dimension de l'espace dans lequel « vivent » les variables devient alors si grand que les données paraissent éparses et éloignées. Il devient alors difficile de dégager des généralités, et de nombreux algorithmes statistiques classiques sont mis à mal par un tel passage à l'échelle. 28

médecine stratifiée approche thérapeutique où l'objectif est de sélectionner les patients auxquels administrer un traitement en fonction d'un marqueur prédictif, afin de ne traiter que la sous-population susceptible de recevoir un bénéfice du traitement. 5

microenvironnement tumoral environnement (cellules, vaisseaux sanguins, molécules) autour de la tumeur. 4

oncogènes gènes dont l'expression favorisent la survenue et le développement des tumeurs. 7

profondeur de séquençage nombre de lectures alignées sur le génome de référence lors de la mesure du niveau d'expression des gènes par la technologie RNA-seq. 11

RNA-Seq technologie permettant de mesurer le niveau d'expression des gènes. 8

survie globale temps entre le dignostic et le décès. 5

survie sans progression temps entre le dignostic et l'appariti ond'un nouvel évènement associé au cancer (récidive ou progression loco-régionale, l'apparation de méta-stases, ou décès du patient avec présence de la tumeur ou de méta-stases). 5

temps de censure temps entre le diagnostic et la censure. 17

temps de suivi temps entre le diagnostic et la fin de l'étude (censure ou observation de l'évènement). 17

temps de survie temps entre le diagnostic et l'évènement étudié (décès, récidive, nouvel évènement associé à la tumeur...). 5

transcription copie simple brin d'un segment particulier de l'ADN, un gène, en ARN dans le noyau de la cellule. 6

Résumé

Le cancer constitue la première cause de mortalité prématurée (décès avant 65 ans) en France depuis 2004. Pour un même organe, chaque cancer est unique, et le pronostic personnalisé est donc un aspect important de la prise en charge et du suivi des patients. La baisse des coûts du séquençage des ARN a permis de mesurer à large échelle les profils moléculaires de nombreux échantillons tumoraux. Ainsi, la base de données TCGA fournit les données RNA-seq de tumeurs, des données cliniques (âge, sexe, grade, stade, etc.), et les temps de suivi des patients associés sur plusieurs années (dont la survie du patient, la récidive éventuelle, etc.). De nouvelles découvertes sont donc rendues possibles en terme de biomarqueurs construits à partir de données transcriptomiques, avec des pronostics individualisés. Ces avancées requièrent le développement de méthodes d'analyse de données en grande dimension adaptées à la prise en compte à la fois des données de survie (censurées à droite), des caractéristiques cliniques, et des profils moléculaires des patients. Dans ce contexte, l'objet principal de la thèse consiste à comparer et adapter des méthodologies pour construire des scores de risques pronostiques de la survie ou de la récidive des patients atteints de cancer à partir de données de séquençage et cliniques.

Le modèle de Cox (semi-paramétrique) est largement utilisé pour modéliser ces données de survie, et permet de les relier à des variables explicatives. Les données RNA-seq de TCGA contiennent plus de 20 000 gènes pour seulement quelques centaines de patients. Le nombre p de variables excède alors le nombre n de patients, et l'estimation des paramètres est soumis à la « malédiction de la dimension ». Les deux principales stratégies permettant de remédier à cela sont les méthodes de pénalisation et le pré-filtrage des gènes. Ainsi, le premier objectif de cette thèse est de comparer les méthodes de pénalisations classiques du modèle de Cox (*i.e.* ridge, lasso, elastic net, adaptive elastic net). Pour cela, nous utilisons des données réelles et simulées permettant de contrôler la quantité d'information contenue dans les données transcriptomiques. Ensuite, la deuxième problématique abordée concerne le pré-filtrage univarié des gènes avant l'utilisation d'un modèle de Cox multivarié. Nous proposons une méthodologie permettant d'augmenter la stabilité des gènes sélectionnés, et de choisir les seuils de filtrage en optimisant les prédictions. Enfin, bien que le coût du séquençage (RNA-seq) ait diminué drastiquement au cours de la dernière décennie, il reste trop élevé pour une utilisation routinière en pratique. Dans une dernière partie, nous montrons que la profondeur de séquençage des miARN peut être réduite sans atténuer la qualité des prédictions pour certains cancers de TCGA, mais pas pour d'autres.

Mots-clés : Cancer, Prédiction, Modèle de Cox, Régression pénalisée, Données de survie, RNA-seq.

Abstract

Cancer has been the leading cause of premature mortality (death before the age of 65) in France since 2004. For the same organ, each cancer is unique, and personalized prognosis is therefore an important aspect of patient management and follow-up. The decrease in sequencing costs over the last decade have made it possible to measure the molecular profiles of many tumors on a large scale. Thus, the TCGA database provides RNA-seq data of tumors, clinical data (age, sex, grade, stage, etc.), and follow-up times of associated patients over several years (including patient survival, possible recurrence, etc.). New discoveries are thus made possible in terms of biomarkers built from transcriptomic data, with individualized prognoses. These advances require the development of large-scale data analysis methods adapted to take into account both survival data (right-censored), clinical characteristics, and molecular profiles of patients. In this context, the main goal of the thesis is to compare and adapt methodologies to construct prognostic risk scores for survival or recurrence of patients with cancer from sequencing and clinical data.

The Cox model (semi-parametric) is widely used to model these survival data, and allows linking them to explanatory variables. The RNA-seq data from TCGA contain more than 20,000 genes for only a few hundred patients. The number p of variables then exceeds the number n of patients, and parameters estimation is subject to the « curse of dimensionality ». The two main strategies to overcome this issue are penalty methods and gene pre-filtering. Thus, the first objective of this thesis is to compare the classical penalization methods of Cox's model (*i.e.* ridge, lasso, elastic net, adaptive elastic net). To this end, we use real and simulated data to control the amount of information contained in the transcriptomic data. Then, the second issue addressed concerns the univariate pre-filtering of genes before using a multivariate Cox model. We propose a methodology to increase the stability of the genes selected, and to choose the filtering thresholds by optimizing the predictions. Finally, although the cost of sequencing (RNA-seq) has decreased drastically over the last decade, it remains too high for routine use in practice. In a final section, we show that the sequencing depth of miRNAs can be reduced without degrading the quality of predictions for some TCGA cancers, but not for others.

Keywords : Cancer, Prediction, Cox model, Penalized regression, Survival data, RNA-seq.