Méthodes statistiques innovantes pour identifier des biomarqueurs dans les essais cliniques en oncologie

Shaima Belhechmi

université
PARIS-SACLAY

Méthodes statistiques innovantes pour identifier des biomarqueurs dans les essais cliniques en oncologie

Innovative statistical methods for the identification of biomarkers in oncology

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 570, Santé publique (EDSP)
Spécialité de doctorat : Santé publique - biostatistiques
Unité de recherche : Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, Villejuif, France
Référent : Faculté de médecine

Thèse présentée et soutenue à Paris-Saclay, le 12 juillet 2021, par

Shaima BELHECHMI

Composition du jury :

Pascale TUBERT-BITTER DR, Directrice de Recherche, Inserm U1018	Présidente
Delphine MAUCORT-BOULCH PU-PH, Université Claude Bernard Lyon 1	Rapporteure & Examinatrice
Thomas FILLERON Phd-HDR, Institut Universitaire du Cancer de Toulouse	Rapporteur & Examinateur
Catherine LEGRAND Pr, Université catholique de Louvain, Belgique	Examinatrice
Boris P. HEJBLUM MCU, Université de Bordeaux	Examinateur
Stefan MICHIELS Phd-HDR, Responsable équipe Oncostat, Université Paris Saclay, INSERM U1018 CESP-OncoStat	Directeur de thèse
Federico ROTOLO PhD, Innate Pharma, Marseille	Co-encadrant

Thèse de doctorat

NNT : 2021UPASR010

Remerciements

Je remercie les membres de mon jury de thèse qui m'ont fait l'honneur d'accepter d'examiner avec attention mon travail. Le Pr Delphine Maucort-Boulch et le Dr Thomas Filleron pour leur relecture attentive. Le Dr Pascale Tubert-Bitter, le Pr Catherine Legrand et le Dr Boris P. Hejblum pour le temps cansacré à revoir mon travail.

Toute ma reconnaissance s'adresse à mon directeur de thèse, Dr Stefan Michiels, et mon co-encadrant, Dr Federico Rotolo, pour m'avoir guidée tout au long de ces années et d'avoir partagé leur expertise. Travailler sous leur direction m'a permis d'entretenir ma curiosité et mes réflexions. A leur confiance, leur rigueur, leur patience, et leur encouragements. Je leur dois mon investissement dans ce travail.

Mes remerciement vont obligatoirement au Gwénaël Le Teuff pour ses conseils, ses encouragements et son soutien inestimables.

Un grand merci au Dr Sarah Flora Jonas pour m'avoir épaulée même dans les moments difficiles.

Enfin, je remercie Mme Zayane Amel qui m'a soutenu pour que je donne le meilleur de moi même.

À la mémoire de ma grand-mère

À mes parents, ma soeur, mes frères et ma petite cousine Maya

Productions scientifiques

Articles

— **Publiés**

Belhechmi S, Michiels S, Paoletti X, et Rotolo F. (**2019**). An alternative trial-level measure for evaluating failure-time surrogate endpoints based on prediction error. *Contemporary clinical trials communications*, 15. https://doi.org/10.1016/j.conctc.2019.100402.

Belhechmi S, De Bin R, Rotolo F, et Michiels S. (**2020**). Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models. *BMC bioinformatics*, 21, 1-20. https://doi.org/10.1186/s12859-020-03618-y.

— **En cours d'écriture**

Belhechmi S, De Bin R, Rotolo F, et Michiels S. Favoring the hierarchy constraint of interactions in penalized survival models of randomized trials.

Communications

— **Orales**

ISCB41 2020, Conférence virtuelle, Cracovie Pologne. Belhechmi S, Rotolo F, et Michiels S. Favoring the hierarchy constraint of interactions in penalized survival models of randomized trials.

ISCB40 2019, Louvain, Belgique. Belhechmi S, De Bin R, Rotolo F, et Michiels S. Accounting for pathways or grouped biomarkers in the high-dimensional penalized Cox model.

EpiClin 2019, Toulouse, France. Belhechmi S, Rotolo F, et Michiels S. Prise en compte des groupes de biomarqueurs ou des voies biologiques dans les modèles de Cox pénalisés de haute dimension.

EpiClin 2018, Nice, France. Belhechmi S, Michiels S, et Rotolo F. Nouvelles mesures de distance pour valider un critère de substitution de type

survie.

SFB : La Journée des Jeunes Chercheurs 2018, Paris, France. Belhechmi S, Rotolo F, et Michiels S. Développement de modèles de survie pronostiques de grande dimension en considérant des groupes de biomarqueurs : application en génétique humaine.

— **Affichées**

IBC 2018, Barcelone, Espagne. Belhechmi S, Rotolo F, et Michiels S. Accounting for pathways or grouped biomarkers in the development of high-dimensional prognostic survival models.

Table des matières

Remerciements ii

Productions scientifiques iii

Table des matières vi

Liste des figures ix

Liste des tableaux xii

1 Introduction générale 1
1.1 Les essais cliniques en oncologie 2
1.2 Les biomarqueurs 8
1.3 Objectifs de la thèse 14
1.4 Références 17

2 Validation d'un critère de substitution 21
2.1 Généralités 23
2.2 État de l'art 25
2.3 Exemple de motivation : méta-analyse GASTRIC 35
2.4 Méthode proposée 39
2.5 Étude de simulation 42
2.6 Application : méta-analyse GASTRIC 48
2.7 Conclusion 49
2.8 Références 53

3 Prise en compte de groupes de variables dans les modèles de régression pénalisés 57
3.1 Introduction 59
3.2 Méthodes statistiques en grande dimension 60
3.3 Approches proposées 65
3.4 Méthodes alternatives 73

3.5 Comparaison des méthodes étudiées 77
3.6 Application : cancer du sein précoce 96
3.7 Discussion 101
3.8 Conclusion 104
3.9 Références 105

4 Favoriser la contrainte hiérarchique des interactions biomarqueurs-traitement 111
4.1 Introduction 113
4.2 Modèle avec interactions sous la contrainte hiérarchique 114
4.3 Méthode proposée 116
4.4 Méthodes alternatives 121
4.5 Étude de simulation 124
4.6 Application : cancer du sein précoce 139
4.7 Conclusion 143
4.8 Références 146

5 Conclusion générale et perspectives 151
5.1 Références 158

Annexes 159

A Résultats complémentaires pour la validation d'un critère de substitution 161

B Résultats complémentaires pour la sélection des biomarqueurs pronostiques dans le contexte de grande dimension 169

Table des figures

1.1 *Distinction entre les biomarqueurs pronostiques et prédictifs* . . . 13

2.1 *Association entre les effets du traitement sur la survie globale et la survie sans progression pour la méta-analyse GASTRIC du cancer digestif avancé* . . . 37
2.2 *La validation croisée d'un contre tous (leave-one-out cross-validation) pour la méta-analyse GASTRIC du cancer digestif avancé* . . . 38
2.3 *Boxplots de distances moyennes pondérées selon les différents scénarios avec les données générées avec la copule de Clayton ajustée* 47

3.1 *Exemple de validation croisée à 5 blocs* . . . 63
3.2 *Résultats de la simulation (a) au niveau des biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs* . . . 88
3.3 *Résultats de la simulation (a) au niveau des biomarqueurs. Boxplots du nombre de biomarqueurs sélectionnés dans tous les scénarios* . . . 89
3.4 *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des groupes de biomarqueurs dans les scénarios alternatifs* . . . 92
3.5 *Les biomarqueurs identifiés par les méthodes sur 500 échantillons d'apprentissage des donnés de l'expression de gènes du cancer du sein pour la survie sans rechute* . . . 99

4.1 *Capacité de sélection des interactions biomarqueurs-traitement dans les scénarios alternatifs* . . . 133
4.2 *Capacité de sélection des effets propres des biomarqueurs prédictifs (interactions) dans les scénarios alternatifs* . . . 137

4.3 *Les effets propres et les interactions gènes-traitement identifiés par les méthodes avec les données de l'expression de gènes du cancer du sein pour la survie sans récidive à distance* 142

A.1 *Boxplots de distances moyennes pondérées selon les différents scénarios avec les données générées avec la copule de Clayton non ajustée* 164
A.2 Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées . . . 165
A.3 Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées . . . 166
A.4 Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées . . . 167

B.1 *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Boxplots du nombre des groupes de biomarqueurs sélectionnés dans tous les scénarios* 173
B.2 *Résultats de la simulation (b) au niveau des biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs* 176
B.3 *Résultats de la simulation (b) au niveau des biomarqueurs. Boxplots du nombre de biomarqueurs sélectionnés dans tous les scénarios* 177
B.4 *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs* 178
B.5 *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Boxplots du nombre des groupes de biomarqueurs sélectionnés dans tous les scénarios* 179
B.6 *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR* 184
B.7 *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR* 185
B.8 *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$* 186
B.9 *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR* 187

B.10 *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR* 188
B.11 *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$* 189
B.12 *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR* 190
B.13 *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR* 191
B.14 *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$* 192
B.15 *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR* 193
B.16 *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR* 194
B.17 *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$* 195

Liste des tableaux

2.1 *Scénarios de l'étude de simulation* 43
2.2 *Distances moyennes pondérées ($\overline{diff}(w)$), erreurs types empiriques (ETE) et erreurs types moyennes (ETM) avec les données générées avec la copule de Clayton ajustée* 46
2.3 *Distances moyennes pondérées entre l'effet observé et l'effet prédit du traitement sur la survie globale pour la méta-analyse GASTRIC du cancer digestif avancé* 48

3.1 La classification des biomarqueurs 82
3.2 *Résultats de la simulation (a). La moyenne et l'erreur standard empirique (ESE) de l'AUC à 5 ans* 95
3.3 *La moyenne et l'erreur type empirique (ETE) de l'AUC à 5 ans pour la prédiction de la survie sans rechute obtenues à partir de 500 échantillons de validation des donnés de l'expression de gènes du cancer du sein* 100

4.1 Scénarios de l'étude de simulation 126
4.2 Capacité de sélection des interactions biomarqueurs-traitement dans tous les scénarios 132
4.3 Capacité de sélection des effets propres des biomarqueurs prédictifs (interactions) dans tous les scénarios 136
4.4 La concordance pour bénéfice et (l'erreur type empirique) des différentes méthodes dans les scénarios alternatifs 138

A.1 Exemple de critères de substitution acceptés par la FDA (*Food and Drug Administration*) 162
A.2 *Distances moyennes pondérées ($\overline{diff}(w_i)$), erreurs types empiriques (ETE) et erreurs types moyennes (ETM) avec les données générées avec la copule de Clayton non ajustée* 163

B.1 *Conception de l'étude de simulation* 170

B.2 *Résultats de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)* 171
B.3 *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Résultats de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de groupes de biomarqueurs sélectionnés)* 172
B.4 *Résultats de la simulation (b) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)* 174
B.5 *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Taux de fausses découvertes / taux de faux négatifs et (le nombre moyen de groupes de biomarqueurs sélectionnés)* 175
B.6 *Résultats de la simulation (b). La moyenne et l'erreur standard empirique (ESE) de l'AUC à 5 ans* 180
B.7 *Résultats analyse de sensibilité au niveau des biomarqueurs pour le choix des paramètres de réglage des méthodes gel et SGL. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)* 181
B.8 Résultats analyse de sensibilité au niveau des groupes de biomarqueurs pour le choix des paramètres de réglage des méthodes gel et SGL. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de groupes de biomarqueurs sélectionnés) 182
B.9 *Résultats de l'étude de sensibilité avec $n = 50$ patients de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / Taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)* 183

Chapitre 1

Introduction générale

Sommaire

1.1 Les essais cliniques en oncologie 2

1.1.1 Quelques enjeux des essais cliniques en oncologie . . 2

1.1.2 Les fonctions liées à la survie 4

1.2 Les biomarqueurs 8

1.2.1 Biomarqueurs de substitution 9

1.2.2 Biomarqueurs pronostiques 10

1.2.3 Biomarqueurs prédictifs 11

1.3 Objectifs de la thèse 14

1.4 Références 17

1.1 Les essais cliniques en oncologie

Les essais cliniques constituent le fondement de la recherche clinique en oncologie et jouent un rôle essentiel dans le développement thérapeutique et la prise de décision. Ils sont définis comme des recherches scientifiques qui évaluent l'efficacité et la sécurité d'une nouvelle approche thérapeutique chez un groupe défini de patients (Kelly et Halabi [2010]). Grâce aux essais cliniques, on peut modifier la pratique médicale, améliorer la qualité de vie des patients ainsi que la survie globale ou la survie sans récidive, etc.
De nombreux essais cliniques sur le cancer impliquent le suivi de patients pendant une longue période de quelques mois à plusieurs années. Le critère de jugement principal dans ces études peut être la survenue du décès, d'une rechute, d'une rémission, d'effets secondaires du traitement ou le développement d'une nouvelle maladie. Au cours de la période de suivi, souvent les données de survie des patients sont recueillies et évaluées. La variable d'intérêt de ces données est le temps qui s'écoule jusqu'à ce qu'un événement se produise (*time-to-event*) c'est-à-dire le temps entre l'inclusion du patient dans l'étude et un évènement ultérieur comme le décès (Singh et Mukhopadhyay [2011]).

1.1.1 Quelques enjeux des essais cliniques en oncologie

Bien que les essais cliniques sont d'une priorité élevée dans la recherche médicale car ils sont une étape nécessaire pour évaluer une nouvelle approche thérapeutique, ils présentent plusieurs enjeux majeurs.
Les essais cliniques randomisés (ECR) évaluant l'efficacité d'un nouveau traitement requièrent une taille d'échantillon de patients suffisamment grande afin de détecter un bénéfice du traitement, s'il existe, en garantissant une puissance statistique suffisante. La taille de l'échantillon doit être augmentée dans les cas

où la probabilité que les patients subissent l'évènement dans un délai compatible avec la durée de suivi planifiée soit faible. En effet, l'analyse de données de survie d'un essai clinique randomisé cherche à rejeter l'hypothèse nulle, à savoir le traitement expérimental a le même effet que le contrôle. À cette fin, il faut que le nombre d'évènements soit élevé pour avoir la certitude qu'une différence clinique entre les deux bras de traitement n'est pas due au hasard. Il s'ensuit que les essais cliniques randomisés peuvent nécessiter une longue période de suivi et des coûts élevés de recherche et de développement. C'est la raison pour laquelle plusieurs essais cliniques cherchent à utiliser des critères de substitution à la place du critère de jugement principal dans le but de réduire le temps nécessaire pour évaluer un nouveau traitement. La définition d'un critère de substitution est présentée dans la section 1.2.1.

Outre cela, de nouveaux défis se sont présentés dans l'ère de la médecine de précision. Jameson et Longo [2015] ont défini la médecine de précision, aussi connue sous le nom de médecine personnalisée ou individualisée, comme "des traitements ciblés aux besoins de chaque patient sur la base de caractéristiques génétiques, biomarqueurs, phénotypiques ou psychosociales qui distinguent un patient donné d'autres patients ayant des profils cliniques similaires". L'idée fondamentale est que l'hétérogénéité dans les caractéristiques des patients implique la nécessité d'individualiser la thérapie (Kravitz et al. [2004]).
La médecine de précision cherche à améliorer les résultats de la pratique médicale en créant des stratégies thérapeutiques adaptées aux caractéristiques individuelles des patients (Mirnezami et al. [2012]). En d'autres termes, elle permet de discriminer les patients selon leur niveau de risque (exemple bas ou haut) et d'identifier des sous-groupes de patients selon leurs caractéristiques biologiques afin d'attribuer le traitement à ceux qui sont susceptibles d'en bé-

néficier. Les récentes avancées des technologies génomiques à haut débit ont conduit à une croissance rapide et une disponibilité plus facile d'informations moléculaires potentiellement pertinentes. Ces caractéristiques biologiques appelées aussi biomarqueurs (présentés dans la section suivante) peuvent aider à améliorer la décision thérapeutique ou à développer de nouveaux traitements ciblés. Toutefois, ce progrès s'accompagne de nouveaux défis tel que l'emploi d'outils statistiques adaptés aux données de grande dimension.

La particularité des données de survie est qu'elles sont soumises au mécanisme de censure, c'est-à-dire qu'il est possible de ne pas observer l'événement d'intérêt pour chaque patient pendant la période de suivi de l'étude. Dans ce cas, on parle de censure à droite, qui est le plus souvent rencontrée dans les essais cliniques. L'analyse de survie est un ensemble d'approches statistiques conçues pour modéliser ce type de données en tenant compte des censures dans l'estimation des quantités d'intérêt. Ces principaux objectifs comprennent l'estimation du temps avant la survenue de l'événement d'intérêt, la comparaison du délai de survenue de l'événement dans différents groupes de patients, et l'évaluation de l'influence de certains facteurs sur le risque d'un événement d'intérêt. Les deux principales fonctions d'intérêt lors de la réalisation d'analyses de survie sont la fonction de survie et la fonction de risque. Nous présentons les principales fonctions liées à la survie.

1.1.2 Les fonctions liées à la survie

Fonction de survie

La fonction de survie $S(t)$ indique la proportion estimée de patients qui

n'ont pas encore présenté l'événement d'intérêt à l'instant t. Elle donne la probabilité que la variable aléatoire associée au temps de survie T dépasse le temps spécifié t. Pour un temps t fixé, la fonction de survie $S(t)$ est définie comme suit :

$$S(t) = \mathbb{P}(T > t), \quad t \geqslant 0. \tag{1.1}$$

$S(t)$ est une fonction décroissante comprise entre 1 et 0.

Fonction de répartition

Les différentes valeurs que peut prendre la variable T ont une distribution de probabilité avec une fonction de densité de probabilité $f(t)$. La fonction de répartition de T, $F(t)$, est alors donnée par :

$$F(t) = \int_0^t f(u)\mathrm{d}u = \mathbb{P}(\mathrm{T} \leq t) = 1 - S(t). \tag{1.2}$$

La fonction de répartition $F(t)$ représente la probabilité que l'événement d'intérêt se produise avant ou à l'instant t. $F(t)$ est une fonction non décroissante, monotone et comprise entre 0 et 1.

Fonction de risque instantané

La fonction de risque instantané $h(t)$ appelée aussi taux de risque (*hazard function*) représente la limite de la probabilité que l'évènement survienne dans un petit intervalle de temps sachant qu'il ne s'est pas produit avant :

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt \mid T \geq t)}{dt} = \frac{f(t)}{S(t)} = -ln(S(t))', \tag{1.3}$$

où dt représente un petit intervalle de temps. La fonction $h(t)$, notée aussi $\lambda(t)$, est une fonction positive.

Fonction de risque cumulé

La fonction de risque instantané $H(t)$ (*cumulative hazard function*) est l'intégrale du risque instantané jusqu'à l'instant t :

$$H(t) = \int_0^t h(u)du = -ln(S(t)). \tag{1.4}$$

$H(t)$ résume le risque cumulé que l'événement d'intérêt (par exemple le décès) survienne avant ou au temps t.

Toutes les fonctions présentées ci-dessus sont étroitement liées les unes aux autres. Afin de décrire la loi de la variable associée au temps de survie T, on peut utiliser soit la fonction de survie $S(t)$ soit la fonction de risque instantané $h(t)$. Plusieurs approches existent pour estimer ces fonctions, nous présentons ci-dessous la méthode de régression la plus utilisée pour modéliser les données de survie.

Modèle à risques proportionnels de Cox

Introduit par le statisticien David Cox (Cox [1972]), le modèle de régression de Cox ou modèle à risques proportionnels est considéré comme l'approche semi-paramétrique la plus répandue dans le domaine de la recherche médicale.
L'objectif principal de ce modèle est d'évaluer l'effet des variables X sur la fonction de risque instantané associée à la durée de survie T sans poser d'hypothèse

concernant la forme de distribution de la fonction survie. Le modèle de Cox exprime un effet multiplicatif des variables X sur le risque instantané $h(t)$:

$$h\left(t;X\right)=h_0(t)\exp\left(X\beta\right)=h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j\right), \tag{1.5}$$

où p est le nombre de variables, β est un vecteur $p\times 1$ de paramètres de régression inconnus et la fonction $h_0(t)$ est le risque de base inconnu donnant le risque en l'absence d'effet des variables ($X=0$) et est le même pour tous les patients. Les coefficients β permettent d'évaluer la force d'association entre les variables explicatives et le critère de jugement principal. Ils sont estimés par maximisation de la fonction de vraisemblance partielle de Cox indépendamment la fonction de risque de base.
Les deux hypothèses du modèle de Cox sont les suivantes :

— L'hypothèse des risques proportionnels stipule que le rapport des risques instantanés (*hazard ratio*, HR) de deux patients i et i' est indépendant du temps.

$$\frac{h\left(t;x_i\right)}{h\left(t;x_{i'}\right)}=\frac{h_0(t)\exp\left(\beta x_i\right)}{h_0(t)\exp\left(\beta x_{i'}\right)}=\exp\left(\beta\left(x_i-x_{i'}\right)\right). \tag{1.6}$$

— L'hypothèse de log-linéarité indique que la relation entre le logarithme du risque instantané et chaque variable est linéaire.

$$ln\left(h\left(t;x_i\right)\right)=ln\left(h_0(t)\right)+\beta x_i. \tag{1.7}$$

En outre, le modèle à risques proportionnels de Cox permet d'évaluer l'effet d'un traitement. Si on suppose Z l'indicateur du bras de traitement ($Z=1$ pour le bras expérimental et $Z=0$ pour le bras contrôle), le rapport de risque HR s'écrit alors :

$$HR = \frac{h(t;Z=1)}{h(t;Z=0)} = \frac{h_0(t)\exp(\beta)}{h_0(t)} = \exp(\beta). \tag{1.8}$$

L'interprétation des valeurs de HR est la suivante :

- $HR = 1 \Leftrightarrow \beta = 0$: Le risque est identique pour les deux bras de traitement
- $HR > 1 \Leftrightarrow \beta > 0$: Le risque est plus élevé pour le bras expérimental
- $HR < 1 \Leftrightarrow \beta < 0$: Le risque est plus faible pour le bras expérimental

De la même manière, le HR peut être calculé pour deux patients qui différent uniquement d'une unité de la variable X_j comme $HR = \exp(\beta_j)$. Si sa valeur est supérieure à 1 alors la variable X_j est considérée comme un facteur de mauvais pronostic et inversement si le $HR < 1$ alors il s'agit d'un facteur de bon pronostic.

1.2 Les biomarqueurs

La définition d'un biomarqueur, ou marqueur biologique, d'après l'institut américain de la santé (NIH) est "une caractéristique mesurée et évaluée objectivement (précision et reproductibilité) comme un indicateur des processus biologiques normaux, des processus pathogènes ou des réponses pharmacologiques à une intervention thérapeutique" (B. D. W. Group et al. [2001]). Il désigne une vaste catégorie d'indicateurs biologiques qui peuvent être mesurés de nombreuses façons, par exemple par biopsie ou par imagerie. Les biomarqueurs peuvent être physiologiques, moléculaires, cellulaires, d'imagerie ou toute autre mesure du patient ou de la tumeur. Buyse et Michiels (Buyse et Michiels [2010]) ont illustré d'une manière explicite et exhaustive les différentes classes de biomarqueurs comme suit.

1.2.1 Biomarqueurs de substitution

Un biomarqueur de substitution, communément appelé critère de substitution, est un critère intermédiaire mesuré précocement qui peut remplacer un critère de jugement principal dans le but d'évaluer l'efficacité d'un traitement expérimental dans un essai clinique randomisé. L'avantage de l'utilisation d'un critère de substitution est qu'il permet de réduire la durée et les coûts de l'essai.
Néanmoins, un biomarqueur doit satisfaire certaines conditions pour être considéré comme un substitut valide. La littérature statistique fournit plusieurs méthodes de validation d'un critère de substitution. Les premières méthodes développées sont basées sur des conditions nécessaires et non suffisantes, comme les critères de Pentice (Prentice [1989]) dont l'un des principaux est que l'effet du traitement sur le critère de jugement principal soit entièrement pris en compte par le critère de substitution. Une autre approche plus récente (Buyse et al. [2000] , Burzykowski et al. [2006]) stipule que la validation d'un substitut doit être sur deux niveaux d'association entre les deux critères :

— L'association au niveau individuel, mesurée à l'aide de données provenant de patients individuels : si elle est forte, cela signifie que le critère de substitution candidat peut estimer le critère de jugement principal au niveau individuel.
— L'association au niveau de l'étude, estimée sur la base de données de plusieurs essais thérapeutiques : si elle est forte, il signifie que l'effet du traitement sur le substitut permet d'estimer l'effet du traitement sur le critère de jugement principal.

Cette approche de validation exige la disponibilité des données individuelles de patients (souvent plus que 1000 patients) issues de plusieurs essais cliniques

randomisés concernant les mêmes questions cliniques spécifiques sur de nouvelles thérapies ou d'un grand essai avec plusieurs unités d'analyse (souvent plus que 10 unités).

En oncologie, les critères de type survie (temps écoulé jusqu'à l'événement) sont couramment utilisés comme un substitut de la survie globale : par exemple, la survie sans maladie (DFS) dans le cancer du côlon adjuvant (SARGENT et al. [2005]) ou la survie sans progression (SSP) dans les études sur la chimiothérapie et la radiothérapie pour des patients atteints d'un cancer du poumon localement avancé (MAUGUEN et al. [2013]).

Pour mieux comprendre les deux autres classes de biomarqueurs (pronostiques et prédictifs), nous présentons la figure 1.1 avec un exemple de courbes de survie en fonction du statut du biomarqueur (positif + ou négatif -) et le traitement reçu (expérimental ou contrôle).

1.2.2 Biomarqueurs pronostiques

Un biomarqueur pronostique est une caractéristique du patient qui renseigne sur la probabilité de survenue de l'événement d'intérêt (par exemple, le décès) indépendamment du traitement ou pour un traitement de référence. L'expression ou non du biomarqueur purement pronostique peut permettre de sélectionner les patients à bas ou à haut risque qui nécessitent un traitement plus agressif ou une surveillance plus intense, mais ne permet pas de prédire directement la réponse à un traitement. Autrement dit, le résultat clinique des patients est corrélé avec le statut du biomarqueur. La figure 1.1 (A) montre un exemple d'un biomarqueur purement pronostique, où le résultat clinique des patients tend à être meilleur lorsque le biomarqueur est positif que lorsqu'il est négatif.

Des exemples de biomarqueurs pronostiques en oncologie sont : les récepteurs hormonaux ($HER2/neu$ et ER) qui sont souvent évalués chez les patientes atteintes d'un cancer du sein, soit au moment du diagnostic soit au moment de la récidive (HARRIS et al. [2007]) ; les délétions du chromosome 17p et les mutations du gène TP53 qui sont associées à un mauvais pronostic chez les patients atteints de leucémie lymphocytaire chronique (SHANAFELT et al. [2006] et GONZALEZ et al. [2011]). On peut trouver également des signatures pronostiques regroupant plusieurs biomarqueurs, comme la signature génomique MammaPrint qui porte sur l'expression de 70 gènes dans le cancer du sein (VAN'T VEER et al. [2002]). Notez que, la validation statistique d'un facteur pronostique nécessite au moins une étude de cas-témoin ou une étude de cohorte (souvent plus que 100 patients).

1.2.3 Biomarqueurs prédictifs

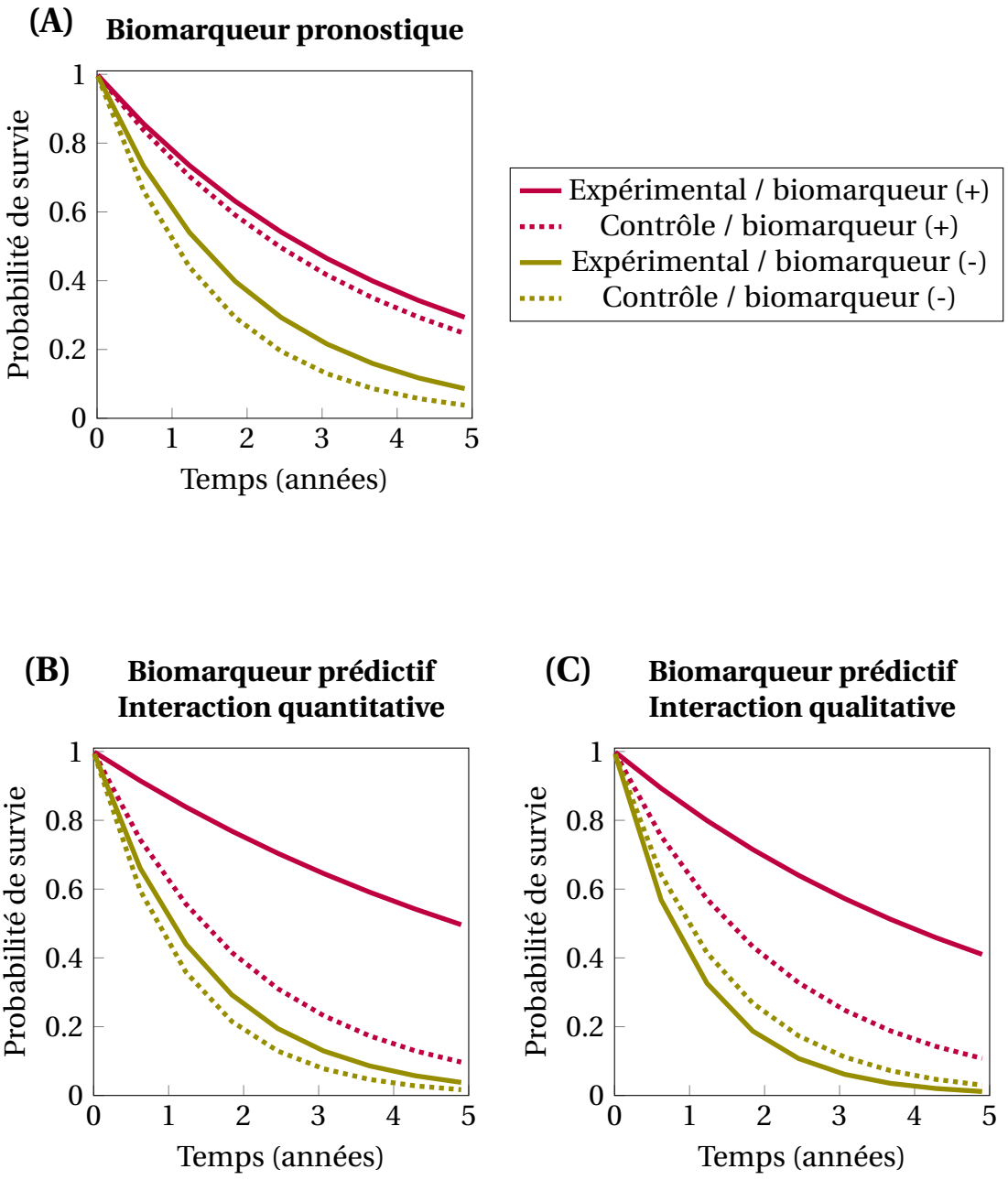
Un biomarqueur est considéré comme prédictif s'il modifie la magnitude de l'effet d'un traitement particulier sur le résultat clinique, d'où l'appellation "modificateur de l'effet du traitement". Les biomarqueurs prédictifs peuvent prédire la différence de résultat clinique, si elle existe, entre deux bras de traitement. Cette interaction avec l'effet du traitement permet de sélectionner le sous-groupe de patients susceptibles de bénéficier du traitement et d'en éviter les effets délétères pour les autres. Pour démontrer l'effet prédictif d'un biomarqueur, il est nécessaire de disposer de données avec un grand nombre de patients (souvent plus que 500 patients) dans un essai clinique randomisé.
La figure 1.1 indique que pour les patients ayant le statut de biomarqueur positif le traitement expérimental est bénéfique, alors que pour les patients ayant un statut négatif, si l'interaction est quantitative (B), le bénéfice du traitement est moins important ; à l'inverse, si l'interaction est qualitative (C), le traite-

ment a un effet délétère.

Bien qu'il soit important de comprendre ces différentes classes de biomarqueurs, il convient de noter qu'un biomarqueur peut être à la fois un facteur pronostique et prédictif. En effet, un biomarqueur prédictif peut être vu comme un biomarqueur ayant un effet pronostique différent dans chaque bras de traitement. C'est le cas, par exemple, chez les patientes atteintes d'un cancer du sein pour lesquelles le statut ER positif est un facteur pronostique favorable (Harris et al. [2007]) et également un facteur prédictif de l'efficacité des thérapies endocriniennes (E. B. C. T. C. Group et al. [2005]). De même, la surexpression du gène $HER2/neu$ est un facteur de mauvais pronostic chez les femmes diagnostiquées avec un cancer du sein précoce, mais constitue un facteur prédictif positif chez les femmes recevant des thérapies ciblées anti-$HER2/neu$, telles que le trastuzumab (Slamon et al. [2001]).

Cependant, la sélection de biomarqueurs pronostiques ou prédictifs dans les situations de grande dimension (identification des gènes pertinents) nécessite l'utilisation de méthodes statistiques appropriées telles que la régression pénalisée.

FIGURE 1.1 : *Distinction entre les biomarqueurs pronostiques et prédictifs*



1.3 Objectifs de la thèse

Les biomarqueurs suscitent un grand intérêt dans la littérature médicale et épidémiologique. En particulier, ils aident à améliorer la pratique clinique et le processus de développement de nouvelles thérapies ainsi qu'à réduire le temps nécessaire à la réalisation des essais cliniques. Les rôles des marqueurs biologiques sont différents les uns des autres, ce qui implique des contraintes statistiques différentes pour les identifier et les valider.

En oncologie, le critère de jugement principal est souvent la survie globale. La validation d'un critère de substitution pour la survie globale doit s'effectuer au niveau individuel et au niveau de l'étude. Le coefficient de détermination R^2 est une mesure de l'association entre les effets traitement sur les deux critères. Sa valeur, si elle est élevée, permet d'assurer la validation au niveau de l'étude. Néanmoins, ce coefficient est peu robuste et très dépendant de l'erreur d'estimation des effets du traitement.
Les données d'expression génique de grande dimension sont collectées rapidement et à faible coût auprès des patients, elles sont caractérisées par un grand nombre de variables (p) et une petite taille d'échantillon (n) ($p \gg n$). L'utilisation d'une méthode statistique non appropriée pour traiter un grand nombre de biomarqueurs engendre un grand nombre de faux positifs. De plus, la plupart des méthodes de sélection de biomarqueurs se concentrent sur des données homogènes, et ne prennent pas en compte les informations disponibles sur leur rôle biologique, par exemple, leur appartenance à des voies biologiques (*pathways*).
D'autre part, l'interaction entre les biomarqueurs et le traitement présente un autre défi. En termes statistiques, la non prise en compte de ces modificateurs

de l'effet du traitement peut conduire à des erreurs dans la conception et l'analyse des essais cliniques. Dans le premier cas, le risque est une mauvaise allocation du traitement aux patients présentant une certaine mutation et dans le second cas, le risque est une mauvaise spécification du modèle de régression (Betensky et al. [2002], Buyse et Michiels [2013]).

Dans le premier axe de ce travail de recherche, nous nous sommes concentrés sur la validation d'un critère de substitution. Nous avons proposé une méthode alternative au coefficient de régression R^2 pour valider un critère de substitution de type survie, basée sur l'erreur de prédiction. Nous avons effectué une étude de simulations et une application sur la méta-analyse GASTRIC évaluant le bénéfice potentiel des différents schémas de chimiothérapies sur la survie globale et la survie sans progression (1^e axe, chapitre 2).
Pour faire la sélection de biomarqueurs pronostiques en tenant compte de leur appartenance à des voies biologiques, nous avons proposé différentes stratégies de pondérations pour la méthode de pénalisation appelée lasso adaptatif qui permettent la sélection à la fois des groupes pertinents et des biomarqueurs à l'intérieur des groupes sélectionnés. Dans un contexte de grande dimension, nous avons comparé l'approche proposée avec d'autres méthodes sous différents scénarios et nous les avons appliquées sur des données réelles d'expression génique mesurées chez des patientes atteintes d'un cancer du sein et traitées par une chimiothérapie adjuvante à base d'anthracyclines (2^e axe, chapitre 3).
Le dernier axe de ce travail de thèse porte sur la prise en compte de la contrainte hiérarchique d'interaction biomarqueurs-traitement lors de la sélection des biomarqueurs prédictifs. Cette contrainte doit permettre de sélectionner des effets propres des biomarqueurs correspondant à des interactions sélection-

nées par le modèle final. Nous avons proposé une stratégie de pondération pour le lasso adaptatif permettant de favoriser cette contrainte hiérarchique. Afin de la comparer avec d'autres méthodes, nous avons mené une étude de simulation et une application sur un jeu de données réelles évaluant l'effet du trastuzumab adjuvant chez des patientes atteintes d'un cancer du sein précoce (3^{e} axe, chapitre 4).

1.4 Références

KELLY, W. & HALABI, S. (2010). *Oncology Clinical Trials : Successful Design, Conduct, and Analysis*. New York : Demos Medical Publishing. 2.

SINGH, R. & MUKHOPADHYAY, K. (2011). Survival analysis in clinical trials : Basics and must know areas. *Perspectives in clinical research*, *2*(4), 145. https://doi.org/10.4103/2229-3485.86872. 2

JAMESON, J. L. & LONGO, D. L. (2015). Precision medicinepersonalized, problematic, and promising. *Obstetrical & gynecological survey*, *70*(10), 612-614. 3.

KRAVITZ, R. L., DUAN, N. & BRASLOW, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, *82*(4), 661-687. https://doi.org/10.1111/j.0887-378X.2004.00327.x. 3

MIRNEZAMI, R., NICHOLSON, J. & DARZI, A. (2012). Preparing for precision medicine. *N Engl J Med*, *366*(6), 489-491. https://doi.org/10.1056/NEJMsb1503104. 3

COX, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, *34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. 6

GROUP, B. D. W., ATKINSON JR, A. J., COLBURN, W. A., DEGRUTTOLA, V. G., DEMETS, D. L., DOWNING, G. J., HOTH, D. F., OATES, J. A., PECK, C. C., SCHOOLEY, R. T. et al. (2001). Biomarkers and surrogate endpoints : preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, *69*(3), 89-95. https://doi.org/10.1067/mcp.2001.113989. 8

BUYSE, M. & MICHIELS, S. (2010). Biomarkers and Surrogate End Points in Clinical Trials. *Oncology Clinical Trials : Successful Design, Conduct, and Analysis* (p. 215-225). New York : Demos Medical Publishing. 8.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials : definition and operational criteria. *Statistics in medicine*, *8*(4), 431-440. https://doi.org/10.1002/sim.4780080407. 9

BUYSE, M., MOLENBERGHS, G., BURZYKOWSKI, T., RENARD, D. & GEYS, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, *1*(1), 49-67. 9.

BURZYKOWSKI, T., MOLENBERGHS, G. & BUYSE, M. (2006). *The evaluation of surrogate endpoints*. Springer Science & Business Media. 9.

SARGENT, D. J., WIEAND, H. S., HALLER, D. G., GRAY, R., BENEDETTI, J. K., BUYSE, M., LABIANCA, R., SEITZ, J. F., O'CALLAGHAN, C. J., FRANCINI, G. et al. (2005). Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies : individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology*, *23*(34), 8664-8670. https://doi.org/10.1200/JCO.2005.01.6071. 10

MAUGUEN, A., PIGNON, J.-P., BURDETT, S., DOMERG, C., FISHER, D., PAULUS, R., MANDREKAR, S. J., BELANI, C. P., SHEPHERD, F. A., EISEN, T. et al. (2013). Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer : a re-analysis of meta-analyses of individual patients' data. *The lancet oncology*, *14*(7), 619-626. https://doi.org/10.1016/S1470-2045(13)70158-X. 10

HARRIS, L., FRITSCHE, H., MENNEL, R., NORTON, L., RAVDIN, P., TAUBE, S., SOMERFIELD, M. R., HAYES, D. F. & BAST JR, R. C. (2007). American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of clinical oncology*, *25*(33), 5287-5312. https://doi.org/10.1200/JCO.2007.14.2364. 11, 12

SHANAFELT, T. D., WITZIG, T. E., FINK, S. R., JENKINS, R. B., PATERNOSTER, S. F., SMOLEY, S. A., STOCKERO, K. J., NAST, D. M., FLYNN, H. C., TSCHUMPER,

R. C. et al. (2006). Prospective evaluation of clonal evolution during long-term follow-up of patients with untreated early-stage chronic lymphocytic leukemia. *Journal of Clinical Oncology*, *24*(28), 4634-4641. https://doi.org/10.1200/JCO.2006.06.9492. 11

Gonzalez, D., Martinez, P., Wade, R., Hockley, S., Oscier, D., Matutes, E., Dearden, C. E., Richards, S. M., Catovsky, D. & Morgan, G. J. (2011). Mutational status of the TP53 gene as a predictor of response and survival in patients with chronic lymphocytic leukemia : results from the LRF CLL4 trial. *Journal of Clinical Oncology*, *29*(16), 2223-2229. https://doi.org/10.1200/JCO.2010.32.0838. 11

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530-536. https://doi.org/10.1038/415530a. 11

Group, E. B. C. T. C. et al. (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival : an overview of the randomised trials. *The Lancet*, *365*(9472), 1687-1717. https://doi.org/10.1016/S0140-6736(05)66544-0. 12

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M. et al. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England journal of medicine*, *344*(11), 783-792. https://doi.org/10.1056/NEJM200103153441101. 12

Betensky, R. A., Louis, D. N. & Gregory Cairncross, J. (2002). Influence of unrecognized molecular heterogeneity on randomized clinical trials. *Journal of Clinical Oncology*, *20*(10), 2495-2499. https://doi.org/10.1200/JCO.2002.06.140. 15

BUYSE, M. & MICHIELS, S. (2013). Omics-based clinical trial designs. *Current opinion in oncology*, *25*(3), 289-295. https://doi.org/10.1097/CCO.0b013e32835ff2fe. 15

Chapitre 2

Validation d'un critère de substitution

Sommaire

2.1 Généralités 23

2.2 État de l'art 25

2.2.1 Approche par essai 25

2.2.2 Approche méta-analytique 28

2.2.2.1 Régression linéaire pondérée 29

2.2.2.2 Modélisation jointe des deux critères d'évaluation 30

2.2.3 Validation croisée 34

2.3 Exemple de motivation : méta-analyse GASTRIC 35

2.4 Méthode proposée 39

2.4.1 Mesures de distance pondérée 39

2.4.2 Choix de pondération 41

2.5 Étude de simulation 42

2.5.1 Génération des données 42

2.5.2 Choix des scénarios 42
2.5.3 Résultats 44
2.6 Application : méta-analyse GASTRIC 48
2.7 Conclusion 49
2.8 Références 53

2.1 Généralités

Un essai clinique est défini comme une étude de recherche visant à répondre à des questions spécifiques sur des vaccins ou de nouvelles thérapies ou de nouvelles façons d'utiliser des traitements connus (Clinical Trials.gov). En oncologie, les essais cliniques peuvent être classés en quatre grandes phases (Crowley et Hoering [2012] et Kelly et Halabi [2010]) :

— Les essais de phase I, menés sur un petit groupe de patients, évaluent la dose maximale tolérée et la sécurité d'un traitement expérimental.

— Les essais de phase II, ont pour objectif d'étudier l'efficacité du traitement expérimental dans la réduction des tumeurs et déterminer la dose optimale.

— Les essais de phase III, sont des essais cliniques randomisés (ECR), destinés à comparer le traitement expérimental à un placebo à fin de confirmer sa supériorité.

  Si les résultats sont prometteurs, une demande d'autorisation de mise sur le marché (AMM) est constituée auprès des autorités de la santé : la FDA (*Food and Drug Administration*) pour les États-Unis, EMA (*European Medicines Agency*) pour l'Europe.

— Les essais de phase IV, sont réalisés après qu'un traitement expérimental ait été mis sur le marché en vue de déterminer tout effet secondaire.

L'efficacité d'un nouveau traitement est évaluée à travers différentes mesures cliniques, dont un critère principal et des critères secondaires, qui correspondent aux objectifs thérapeutiques d'un traitement. Le critère de jugement principal (*primary endpoints*) est la variable la plus pertinente permettant de répondre à la question clinique d'une étude de recherche (Kramar et Mathoulin-Pélissier [2011]). En cancérologie, le critère de jugement prin-

cipal des essais cliniques randomisés de phase III est souvent la survie globale (SG) et l'événement d'intérêt peut être le décès. Cependant, l'utilisation de ce critère de jugement principal requiert un nombre important de décès, une longue période de suivi et des coûts de recherche et de développement élevés pour atteindre la puissance statistique nécessaire. Par exemple, dans une étude randomisée de phase III dans le cancer digestif métastasique avancé (Boku et al. [2009]), il a fallu plus de 5 ans pour randomiser au moins 700 patients afin de mettre en évidence un bénéfice de survie globale de 10% avec une puissance de 70%. Ce problème est encore plus important dans le cas des maladies rares ou les pathologies moins avancées pour lesquelles le nombre de décès est relativement faible par rapport au nombre de patients inclus sur le court terme. Il est donc plus judicieux d'utiliser des critères intermédiaires ou de substitution (*surrogate*) à la place d'un critère de jugement principal. En effet, un critère de substitution est un biomarqueur permettant de mesurer précocement les bénéfices cliniques d'un traitement, ce qui constitue un gain en termes de nombre de patients inclus, de durée et de coûts de l'essai. Un critère intermédiaire peut également être intéressant s'il peut être mesuré plus rapidement, de manière plus précise et moins invasive pour les patients par rapport à un critère de jugement principal. En oncologie, les critères intermédiaires fréquemment utilisés comme substitut de la survie globale, sont le temps de survie sans progression (SSP), ou sans rechute (SSR), ou encore sans événement (SSE).

Toutefois, avant d'utiliser un critère intermédiaire, il est nécessaire de le valider comme un critère de substitution au critère de jugement principal. De fait, il faut que les conclusions cliniques soient cohérentes entre les deux critères. A titre d'illustration, la FDA maintient à jour une liste des critères de substitution acceptés qui sert de guide de référence pour les nouveaux essais cliniques (voir

des exemples dans l'annexe A tableau A.1).

2.2 État de l'art

Dans le cadre des essais cliniques randomisés de phase III, nous présentons dans cette section des méthodes de validation des critères de substituions. Deux approches ont été développées, la première est basée sur un seul essai, et la deuxième est une approche méta-analytique. La méta-analyse est une synthèse exhaustive et quantitative de tous les essais thérapeutiques concernant la même question. Avec l'augmentation du nombre de patients inclus elle entraîne une augmentation de la puissance statistique.

2.2.1 Approche par essai

Critères de Prentice

Les premiers travaux effectués pour valider un critère de substitution utilisent l'information apportée par un seul essai clinique randomisé. PRENTICE [1989] (voir aussi KRAMAR et MATHOULIN-PÉLISSIER [2011]) fut le premier à présenter une définition statistique d'un critère de substitution, soit "une variable réponse pour laquelle un test d'hypothèse nulle d'absence de relation à un groupe de traitement est également un test valide de l'hypothèse nulle correspondante basée sur le vrai critère de jugement". Cette définition s'écrit mathématiquement comme suit :

$$f(S \mid Z) = f(S) \iff f(T \mid S) = f(T), \tag{2.1}$$

où les variables T et S correspondent au critère de jugement principal et au critère de substitution, la variable binaire Z correspond au bras de traitement (Z

= 1 pour traitement expérimental et $Z = 0$ pour le contrôle), et $f(.)$ et $f(. \mid .)$ dénotent, respectivement, la distribution de probabilité et la distribution conditionnelle. Prentice a proposé quatre critères opérationnels qui doivent être vérifiés pour qu'un critère candidat soit un critère de substitution :

1. $f(S \mid Z) \neq f(S)$: l'effet du traitement sur le critère de substitution est statistiquement significatif.
2. $f(T \mid Z) \neq f(T)$: l'effet du traitement sur le critère de jugement principal est statistiquement significatif.
3. $f(T \mid S) \neq f(T)$: l'effet pronostique du critère de substitution sur le critère de jugement principal est statistiquement significatif, c'est-à-dire, le critère de substitution est un bon prédicteur du critère de jugement principal.
4. $f(T \mid S, Z) = f(T \mid S)$: après ajustement sur le critère de substitution, l'effet du traitement sur le critère de jugement principal n'est plus statistiquement significatif, c'est-à-dire, le critère de substitution capture complètement l'effet du traitement.

Cependant, ces 4 critères reposent sur des tests d'hypothèse qui sont dépendants des erreurs de type I et de type II. Par conséquent, un manque de puissance peut engendrer des conclusions erronées. De plus, il est difficile voir impossible qu'un critère de substitution capture totalement l'effet du traitement et donc l'hypothèse nulle ne sera jamais 'complètement vraie' et un test statistique ne peut pas valider l'hypothèse nulle.

Proportion expliquée de Freedman

Freedman et al. [1992] ont proposé d'estimer la proportion d'effet du traitement (PE) expliquée par le critère de substitution comme mesure de la validité

d'un substitut potentiel.

$$PE = \frac{\beta - \beta_s}{\beta} = 1 - \frac{\beta_s}{\beta}, \tag{2.2}$$

où β et β_s sont les estimations des effets du traitement sur le critère de jugement principal, sans et avec ajustement sur le critère de substitution, respectivement. Ces estimations peuvent être obtenues par exemple par un modèle de régression de Cox (Cox [2000]) pour les données de survie. Le quatrième critère de Prentice nécessite que $\beta_s = 0$, soit que $PE = 1$.
Buyse et al. [2016] ont souligné les inconvénients de l'utilisation de cette mesure, puisqu'elle n'est pas une valeur stable et n'est pas limitée à l'intervalle [0,1] : elle peut être négative dans certains cas ou supérieure à 1 si β et β_s sont de signe opposé. En outre, la PE nécessite un fort effet du traitement sur le critère de jugement principal et un grand nombre d'observations, ce qui pourrait ne pas être le cas dans la plupart des essais cliniques randomisés et pourrait conduire alors à de larges intervalles de confiance.

Effet relatif et association ajustée

Buyse et Molenberghs [1998] ont proposé deux mesures alternatives à la PE. La première, définie au niveau de la population et appelée effet relatif (RE), est le ratio de l'effet du traitement sur le critère de jugement principal et l'effet sur le critère de substitution. Si on suppose α le paramètre associé à l'effet du traitement sur le critère de substitution,

$$RE = \frac{\beta}{\alpha}. \tag{2.3}$$

La deuxième mesure, dénommée association ajustée γ_Z, est l'association individuelle entre le critère de substitution et le critère de jugement principal, après

ajustement de l'effet du traitement. Ainsi, $PE = \frac{\gamma_Z}{RE}$ est le ratio de l'association ajustée (une mesure au niveau individuel) et de l'effet relatif (une mesure au niveau de l'essai).

Bien que toutes ces propositions soient relativement faciles à calculer, elles sont limitées par le fait qu'elles sont basées sur des données d'un essai unique. Elles ne permettent donc uniquement que de déterminer des associations au niveau individuel. D'où la nécessité de l'utilisation des données provenant de plusieurs essais.

2.2.2 Approche méta-analytique

Plusieurs auteurs, Daniels et Hughes [1997] , Buyse et al. [2000] , et Gail et al. [2000] ont introduit l'approche méta-analytique afin d'utiliser les données individuelles collectées de plusieurs essais cliniques randomisés. Cette approche a été formulée à l'origine pour deux critères continus normalement distribués. Elle a été étendue à un large éventail de types de critères comme les critères binaires, ordinaux, de survie ou encore des critères mesurés longitudinalement (Burzykowski et al. [2006]).

L'approche méta-analytique, basée sur la corrélation entre les deux critères, fournit une prédiction de l'effet du traitement sur le critère de jugement principal à partir de l'effet du traitement sur le critère de substitution. Elle considère deux niveaux de validation le niveau individuel et le niveau de l'étude :

— L'association au niveau individuel signifie que, pour chaque patient, le critère de substitution candidat est corrélé au critère de jugement principal.

— L'association au niveau de l'étude signifie que, pour chaque essai, l'effet du traitement sur le critère de substitution est corrélé avec l'effet sur le critère de jugement principal.

Ainsi, un critère de substitution valide doit être corrélé au critère de jugement principal et doit prédire l'effet du traitement sur ce critère.

2.2.2.1 Régression linéaire pondérée

La validation d'un critère de substitution au deux niveaux peut se faire par une régression linéaire pondérée.

Association au niveau individuel

Afin de mesurer l'association au niveau individuel, on modélise la relation entre le critère de substitution S et le critère de jugement principal T par une régression linéaire des taux de survie pondérée par la taille de chacun des essais. Ainsi, un coefficient de corrélation de rang peut être calculé entre les deux critères.

Association au niveau de l'étude

Pour un essai clinique randomisé comparant le bras expérimental au bras contrôle et portant sur un critère de survie, l'effet du traitement est modélisé par le logarithme du rapport de risque instantané (*Hazard Ratio*) ($ln(HR)$). Le HR est estimé par le modèle semi-paramétrique de Cox. L'association au niveau de l'étude est mesurée par une régression linéaire des effets du traitement sur le critère de substitution ($ln(HR_S)$) et le critère de jugement principal ($ln(HR_T)$) pondérée par la taille respective de chacun des essais. Cette régression permet de calculer le coefficient de détermination R^2 reflétant la force de l'association entre les effets du traitement sur les deux critères. Le coefficient R^2 représente la part de variance du $ln(HR_T)$ expliquée par le modèle de régression linéaire.

Cependant, ce modèle de régression univariée ne prend pas en compte l'in-

certitude des estimations (les erreurs types) des effets du traitement sur les deux critères $ln(\widehat{HR}_T)$ et $ln(\widehat{HR}_S)$. Par conséquent, il est possible de surestimer la précision de la prédiction obtenue par régression linéaire ainsi que le coefficient de détermination R^2, en particulier lorsque les essais cliniques sont de taille modeste. VAN HOUWELINGEN et al. [2002] ont proposé d'intégrer ces erreurs types dans la régression linéaire à l'aide d'un modèle tenant compte des incertitudes liées au processus d'estimation. Cette approche permet à la fois de donner plus de poids aux essais dont les estimations des effets du traitement sont plus précises et d'estimer la précision associée au coefficient de détermination R^2.

2.2.2.2 Modélisation jointe des deux critères d'évaluation

BUYSE et al. [2000] ont suggéré deux approches de modélisation jointe des deux critères d'évaluation pour la validation d'un critère de substitution. Ces approches ont été formulées à l'origine pour des critères continus, normalement distribués. La première approche se base sur un modèle en deux étapes et à effet fixe, et la deuxième combine les deux étapes en utilisant un modèle linéaire à effet mixtes. BURZYKOWSKI et al. [2001] ont proposé d'étendre le modèle en deux étapes (*two-stage model*) pour les critères de type survie en utilisant lors de la première étape un modèle de copule bivarié. Soient T_{ij} et S_{ij} deux variables aléatoires indiquant, respectivement, le critère de jugement principal et le critère de substitution pour le patient j dans l'étude i, ainsi la fonction de survie jointe de (T_{ij}, S_{ij}) s'écrit comme suit :

$$\begin{cases} h_{Sij}(s; Z_{ij}) = h_{Si}(s) \exp\left\{(\alpha_i) Z_{ij}\right\} \\ h_{Tij}(t; Z_{ij}) = h_{Ti}(t) \exp\left\{(\beta_i) Z_{ij}\right\} \\ S_{ij}(s, t; Z_{ij}) = C_\delta\left\{S_{Sij}(s; Z_{ij}), S_{Tij}(t; Z_{ij})\right\}, \ s, t \geq 0, \end{cases} \tag{2.4}$$

avec $h_{Si}(s)$ et $h_{Ti}(s)$ les fonctions de risque de base spécifiques à l'essai, α_i et β_i les effets du traitement sur les deux critères respectifs, et Z l'indicateur du bras de traitement. La fonction copule de Clayton $C_\delta(S_{Sij}(s), S_{Tij}(t))$ (CLAYTON [1978]) est une fonction de distribution bivariée définie de la manière suivante :

$$C_\delta(u, v) = \left(u^{-\delta} + v^{-\delta} - 1\right)^{-1/\delta}, \quad \delta > 0. \tag{2.5}$$

La copule de Clayton décrit l'association entre T_{ij} et S_{ij}. La force de cette association diminue avec la diminution de δ et atteint l'indépendance lorsque $\delta \to 1$. Les fonctions de survie marginales sont basées sur l'hypothèse des risques proportionnels pour chacun des deux critères cliniques. Ainsi la distribution de Weibull est utilisée pour spécifier, d'une façon paramétrique, les fonctions de risque de base.

L'estimation du modèle (2.4) est effectuée en deux étapes. Dans la première étape, les effets du traitement sur les deux critères cliniques sont estimés pour chaque essai par le maximum de la vraisemblance du modèle de copule de Clayton (2.5). Ainsi, un coefficient de corrélation peut être calculé pour valider le critère de substitution au niveau individuel. Dans la deuxième étape, une régression linéaire qui tient compte des erreurs d'estimation à la première étape est effectuée pour estimer la relation entre les effets du traitement sur les deux critères. Le critère de substitution est considéré comme valide au niveau de l'étude s'il est capable de prédire l'effet du traitement sur le critère de jugement principal avec une précision suffisante.

Association au niveau individuel

Le tau (τ) de Kendall (KENDALL [1938]) est un coefficient de corrélation souvent utilisé pour la validation au niveau individuel d'un critère de substitution de type survie. Il calculé en fonction de la copule de la manière suivante :

$$\tau = 4 \int_0^1 \int_0^1 C_\delta(u,v) C_\delta(\mathrm{d}u, \mathrm{d}v) - 1. \tag{2.6}$$

Pour la couple de Clayton, le τ de Kendall peut être dérivé à partir de l'estimation de δ, comme :

$$\tau = \frac{\delta}{\delta + 2}. \tag{2.7}$$

Le τ de Kendall mesure la différence entre la probabilité de concordance et celle de discordance entre T et S pour deux patients données. Il est compris dans l'intervalle [-1,1]. Une valeur de $\tau = 0$ signifie l'absence d'association, les valeurs -1 et $+1$ signifient, respectivement, une parfaite corrélation négative et positive.

Association au niveau de l'étude

L'association au niveau de l'étude est évaluée à partir du coefficient de détermination R^2_{essai}. Ce coefficient de détermination est estimé dans une régression linéaire des effets du traitement qui prend en compte les erreurs d'estimation issues de la première étape pour réduire le biais.
Les estimations des effets du traitement obtenues lors de la première étape sont supposées suivre le modèle mixte suivant :

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \epsilon_{ai} \\ \epsilon_{bi} \end{pmatrix},$$

où les vrais effets du traitement sont normalement distribués avec une matrice de dispersion $\boldsymbol{D}$

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \boldsymbol{D}\right), \quad \boldsymbol{D} = \begin{pmatrix} d_\alpha^2 & d_\alpha d_\beta \rho_{\text{essai}} \\ d_\alpha d_\beta \rho_{\text{essai}} & d_\beta^2 \end{pmatrix},$$

et les erreurs d'estimation ϵ_{ai} et ϵ_{bi} sont normalement distribués avec une moyenne nulle et une matrice de covariance $\boldsymbol{\Omega}_i$

$$\begin{pmatrix} \epsilon_{ai} \\ \epsilon_{bi} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_i\right), \quad \boldsymbol{\Omega}_i = \begin{pmatrix} \omega_{\alpha_i}^2 & \omega_{\alpha_i}\omega_{\beta_i}\rho_{\epsilon_i} \\ \omega_{\alpha_i}\omega_{\beta_i}\rho_{\epsilon_i} & \omega_{\beta_i}^2 \end{pmatrix}.$$

Le coefficient de détermination R^2_{essai} est estimé en fixant les matrices de covariance Ω_i à leurs valeurs estimées à partir du modèle de copule bivarié à la première étape (van Houwelingen et al. [2002]) . L'estimation du R^2_{essai} ajusté par les erreurs d'estimation est définie comme :

$$\hat{R}^2_{essai} = \frac{d_\alpha^2 d_\beta^2 \rho^2_{trial}}{d_\alpha^2 d_\beta^2} = \rho^2_{essai}. \tag{2.8}$$

Le $\hat{R}^2_{essai}$ est compris entre [0,1]. Une valeur de $\hat{R}^2_{essai}$ suffisamment proche de 1 signifie une force d'association élevée entre les effets du traitement sur le critère de substitution et le critère de jugement principal.

En pratique, l'algorithme utilisé pour estimer les effets du traitement $\left(\hat{\alpha}_i, \hat{\beta}_i\right)$ ne converge pas souvent et un modèle non ajusté par les erreurs d'estimation est utilisé avec des matrices de covariance Ω_i fixés à zéro. Toutefois, le coefficient R^2 est une mesure peu robuste, c'est-à-dire très sensible aux valeurs extrêmes et dans certains cas son erreur type est assez large.

D'autre part, la comparaison de l'effet observé du traitement sur le critère de jugement principal ($ln(\widehat{HR}_T)$) et sa prédiction basée sur l'effet observé du

traitement sur le critère de substitution ($ln(\widehat{HR}_S)$) dans la même étude permet de juger si le modèle de prédiction est plus au moins parcimonieux. L'exactitude et la précision du modèle dépend de la force d'association au niveau de l'étude. Cependant, l'utilisation de la même information apportée par un groupe d'études pour à la fois estimer et évaluer le modèle de prédiction pose un problème de sur-ajustement du modèle et de sur-optimisme des conclusions que l'on peut en tirer.

2.2.3 Validation croisée

La validation croisée d'un contre tous (*leave-one-out cross-validation* ou LOOCV) peut être utilisée pour valider un modèle statistique grâce à une information externe au modèle afin d'évaluer sa performance tout en limitant les problèmes de sur-apprentissage. MICHIELS et al. [2009] ont proposé de comparer l'effet observé du traitement sur le critère de jugement principal avec une prédiction indépendante en utilisant la validation croisée d'un contre tous (LOOCV) avec une approche méta-analytique pour valider un critère de substitution.

Pour une méta-analyse avec N essais, cette méthode consiste à exclure un seul essai à la fois, puis, à estimer le modèle en deux étapes (2.4) avec les $N-1$ essais restants, et enfin, à valider ce modèle de prédiction sur l'essai exclu. Cette dernière étape consiste à intégrer l'effet observé du traitement sur le critère de substitution ($ln(\widehat{HR}_S)$) de l'essai exclu, estimé par le modèle de Cox, dans le modèle de prédiction afin de prédire l'effet du traitement sur le critère de jugement principal ($ln(\widetilde{HR}_T)$) de l'étude exclue et calculer un intervalle de prédiction à 95%. Cet effet prédit du traitement ($ln(\widetilde{HR}_T)$) est finalement comparée à l'effet observé du traitement ($ln(\widehat{HR}_T)$) sur le critère de jugement principal calculé aussi avec le modèle semi-paramétrique de Cox.

Pour une méta-analyse de N essais, les étapes de la LOOCV sont les suivantes :

1. Exclure un essai, supposons le k^{e} essai, $(k = 1, \ldots, N)$
2. Ajuster le modèle en deux étapes 2.4 avec les $N - 1$ essais restants $\rightarrow$ obtenir un modèle de prédiction
3. Estimer les effets du traitement sur le critère de substitution $ln(\widehat{HR}_S)$ et le critère de jugement principal $ln(\widehat{HR}_T)$ de l'essai exclu (le k^{e} essai) avec le modèle de Cox
4. Prédire l'effet du traitement sur le critère de jugement principal $ln(\widetilde{HR}_T)$ de l'essai exclu et calculer l'intervalle de prédiction à 95% en intégrant le $ln(\widehat{HR}_S)$ dans le modèle de prédiction construit avec les $N - 1$ essais
5. Répéter les étapes précédentes N fois jusqu'à l'obtention des $ln(\widehat{HR}_T)$ et $ln(\widetilde{HR}_T)$ pour les N essais.

Enfin, l'effet prédit du traitement $ln(\widetilde{HR}_T)$ est comparé à l'effet observé du traitement $ln(\widehat{HR}_T)$ (voir l'exemple de la figure 2.2 dans la section suivante 2.3).

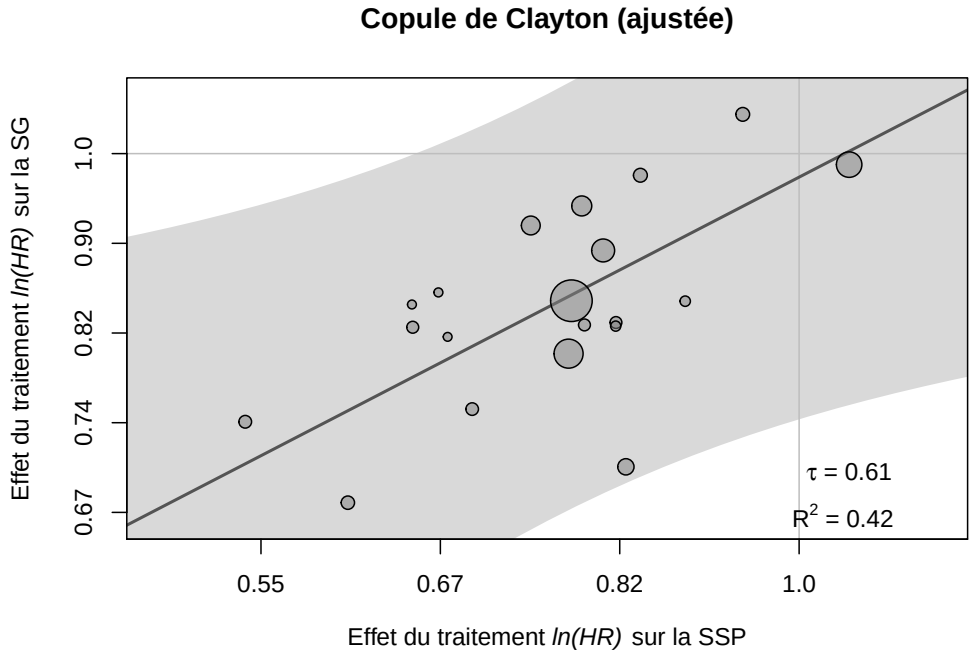
2.3 Exemple de motivation : méta-analyse GASTRIC

La méta-analyse GASTRIC (Group et al. [2013]) d'essais cliniques randomisés basés sur des données individuelles de patients évalue le bénéfice potentiel des différents schémas de chimiothérapies sur la survie globale et la survie sans progression. Cette méta-analyse comprend 4069 patients atteints d'un cancer digestif avancé/récurrent provenant de 20 essais randomisés de chimiothérapie. Le critère clinique principal de jugement était la survie globale et le critère de substitution candidat était la survie sans progression. Dans des travaux antérieurs (Paoletti et al. [2013] et Rotolo, Paoletti, Burzykowski

et al. [2017]), l'association au niveau individuel entre la survie globale (SG) et la survie sans progression (SSP), telle que mesurée par le τ de Kendall était de 0,61 avec un intervalle de confiance à 95% [95% IC : 0,59 -0,62] et de 0,68 [95% IC : 0,68 -0,68] pour la copule de Clayton ajustée et Plackett ajustée respectivement. L'association au niveau de l'étude entre les effets du traitement sur la survie globale et sur la survie sans progression était modérée. Le coefficient de détermination R^2 ajusté par les erreurs d'estimation était, respectivement, de 0,41 [95% IC : 0,15 -1,00] et de 0,61 [95% IC : 0,04 -1,00] pour la copule de Clayton ajustée et Plackett ajustée (voir la figure 2.1 avec la copule de Clayton ajustée). Comme dans de nombreuses applications, la largeur de l'intervalle de confiance suggère qu'il existe une incertitude considérable dans le modèle de prédiction et qu'il faut faire preuve de prudence dans l'interprétation de ces résultats.

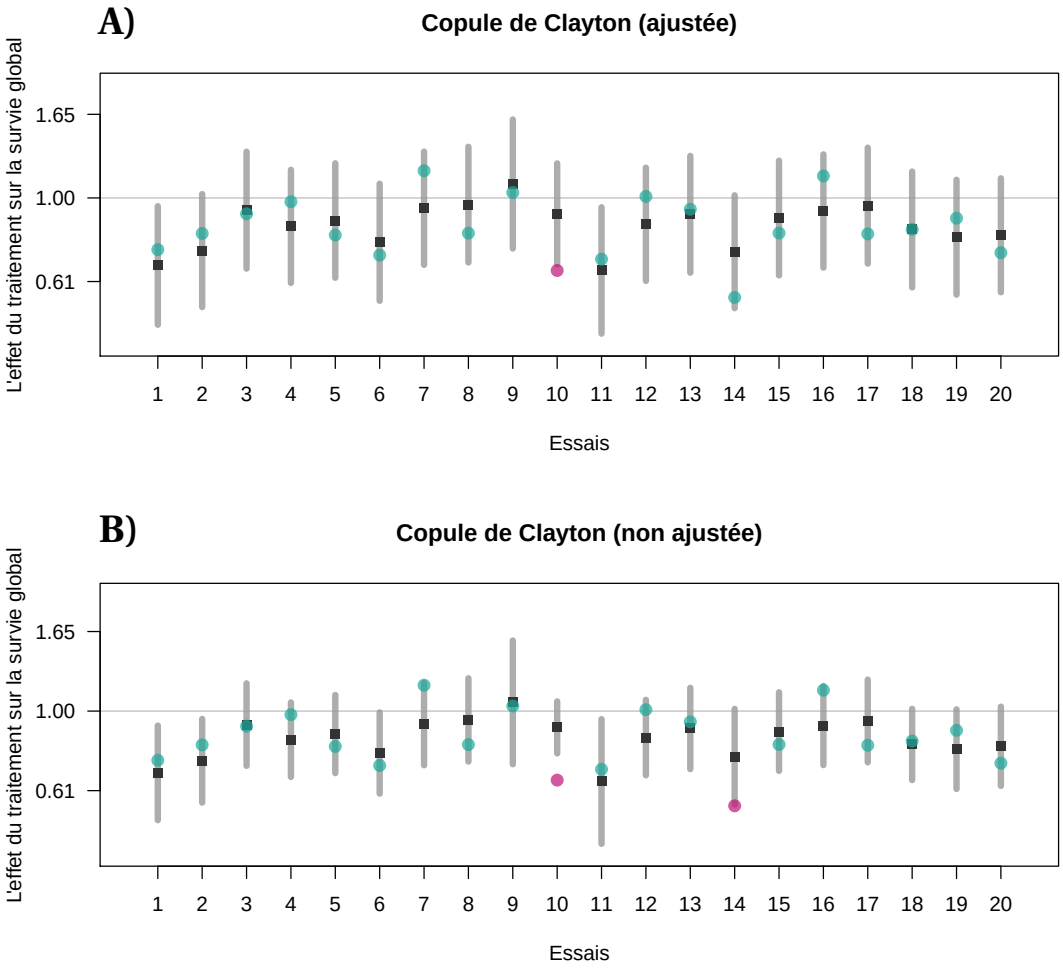
La figure 2.2 présente les résultats de la comparaison de l'effet prédit ($\widetilde{HR}_{SG}$) et observé ($\widehat{HR}_{SG}$) du traitement sur la survie globale avec la validation croisée d'un contre tous (LOOCV) et la copule de Clayton ajustée (A) et non ajustée (B) par les erreurs d'estimation. Les lignes verticales sont les intervalles de prédiction au niveau 95% de l'effet du traitement sur la survie globale basés sur le modèle de prédiction estimé sur 19 essais. Les points en bleu sont les effet observé du traitement $\widehat{HR}_{SG}$ qui se trouves dans l'intervalle de prédiction et ceux qui sont en violet sont en dehors de l'intervalle. Si le critère de substitution (la survie sans progression) est valide au niveau de l'étude, alors il est capable de prédire avec une précision suffisante l'effet du traitement sur la survie globale. Ainsi, on s'attend à ce qu'environ 95% des $\widehat{HR}_{SG}$ se situent à l'intérieur de l'intervalle de prédiction. Dans notre exemple, cela est vrai dans 19 essais sur 20 (95%) pour le modèle de Clayton ajusté et 18 essais sur 20 (90%) pour le modèle de Clayton non ajusté.

FIGURE 2.1 : *Association entre les effets du traitement sur la survie globale et la survie sans progression pour la méta-analyse GASTRIC du cancer digestif avancé*



L'échelle logarithmique a été utilisée pour les deux axes x et y. Les cercles représentent les coordonnées des effets du traitement pour chaque étude, leur taille est proportionnelle au nombre de patients inclus dans chaque étude, la bande grise est la bande de confiance des estimations. SG : survie globale, SSP : survie sans progression

FIGURE 2.2 : *La validation croisée d'un contre tous (leave-one-out cross-validation) pour la méta-analyse GASTRIC du cancer digestif avancé*



Les carrés noirs et les lignes verticales grises sont les valeurs prédites de l'effet du traitement sur la survie globale, avec les intervalles de prédiction à 95%. Les points représentent les effets observés du traitement sur la survie globale (vert = dans l'intervalle de prédiction, magenta = hors de l'intervalle de prédiction)

2.4 Méthode proposée

La comparaison des effets prédits et observés du traitement (figure 2.2), nous a incité à proposer de nouvelles mesures de validation d'un critère de substitution au niveau de l'étude dans le contexte méta-analytique. Ces mesures, alternatives au coefficient de détermination R^2_{essai}, sont basées sur la distance entre la valeur observée de l'effet du traitement sur le critère de jugement principal $\hat{\beta} = ln(\widehat{HR}_T)$ et la valeur prédite $\tilde{\beta} = ln(\widetilde{HR}_T)$. En particulier, nous avons supposé que l'erreur de prédiction en termes de distance entre la valeur prédite et la valeur observée $diff = \left|\hat{\beta} - \tilde{\beta}\right|$ est la mesure la plus importante. Également, nous avons proposé différentes pondérations pour attribuer un poids plus important aux essais avec des intervalles de prédiction plus étroits et d'estimations plus précises que ceux avec des estimations et des prédictions moins précises.

2.4.1 Mesures de distance pondérée

Soit $i = 1, \ldots, N$, l'indice de l'essai clinique d'une méta-analyse, et $\hat{\alpha}_i$ et $\hat{\beta}_i$ les effets du traitement sur le critère de substitution et le critère de jugement principal estimés séparément par les deux modèles à risques proportionnels marginaux 2.4, respectivement. Notons $\tilde{\beta}_i$ la prédiction de $\hat{\beta}_i$ obtenue en injectant l'effet observé sur le critère de substitution $\hat{\alpha}_i$ dans le modèle de prédiction estimé par les $N - 1$ essais (LOOCV). La prédiction de l'effet du traitement sur le critère de jugement principal peut être obtenue de la manière suivante :

$$\tilde{\beta}_i = \tilde{\gamma}_0 + \tilde{\gamma}_1 \hat{\alpha}_i, \tag{2.9}$$

avec

$$\tilde{\gamma}_1 = \frac{(N-1)S_{ab} - \sum\limits_{i' \neq i} \hat{\omega}_{\alpha i'} \hat{\omega}_{\beta i'} \hat{\rho}_{\text{trial}}}{(N-1)S_{aa} - \sum\limits_{i' \neq i} \hat{\omega}^2_{\alpha i'}}, \tag{2.10}$$

et

$$\tilde{\gamma}_0 = \frac{\sum\limits_{i' \neq i} \left(\hat{\beta}_{i'} - \tilde{\gamma}_1 \hat{\alpha}_{i'} \right)}{(N-1)}, \tag{2.11}$$

où S_{aa} et S_{ab} sont la variance et la covariance de l'échantillon des estimations $\hat{\alpha}_i$ et $\hat{\beta}_i$ (Burzykowski et Abrahantes [2005]).
Soit $diff_i$ la différence absolue entre l'effet observé du traitement sur le critère de jugement principal $\hat{\beta}_i$ et sa prédiction $\tilde{\beta}_i$ dans l'essai i :

$$diff_i = \left| \hat{\beta}_i - \tilde{\beta}_i \right|. \tag{2.12}$$

La différence absolue $diff_i$ représente l'erreur de prédiction en termes de distance entre la valeur prédite et la valeur observée (voir la figure 2.2).
Nous avons proposé diverses distances absolue moyennes pondérées $\overline{diff}(w)$ pour tenir compte de la précision des effets prédits et observés du traitement, qui peuvent être différents d'un essai à l'autre, comme suit :

$$\overline{diff}(w) = \frac{\sum_{i=1}^{N} diff_i \times w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} \left| \hat{\beta}_i - \tilde{\beta}_i \right| \times w_i}{\sum_{i=1}^{N} w_i}, \tag{2.13}$$

où w_i, $(i = 1 \ldots N)$, le poids spécifique de l'essai i. Nous avons également défini l'erreur type (*standard erreur (SE)*) de la distance absolue moyenne pondérée de la manière suivante :

$$SE\left(\overline{diff}(w)\right) = \frac{\sum_{i=1}^{N}\left(diff_i - \overline{diff}(w)\right)^2 \times w_i}{\sum_{i=1}^{N} w_i}. \tag{2.14}$$

2.4.2 Choix de pondération

Nous avons proposé six approches de pondération, le poids pour chaque essai de la méta-analyse se présente comme suit :

- $w_i = 1$, pour calculer la distance absolue moyenne non pondérée.
- $w_i = n_i$, pour donner plus d'importance aux essais de plus grande taille.
- $w_i = 1/SE^2(\hat{\beta}_i)$, l'inverse du carré de l'erreur type (*standard erreur (SE)*) de $\hat{\beta}_i$, pour donner plus de poids aux essais avec des estimations plus précises de l'effet observé du traitement.
- $w_i = 1/SE^2(\tilde{\beta}_i)$, l'inverse du carré de l'erreur type de $\tilde{\beta}_i$, pour donner plus de poids aux essais avec des prédictions plus précises de l'effet du traitement.
- $w_i = 1/\left(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i)\right)$, pour tenir compte de la précision des valeurs prédites et observées.
- $w_i = 1/\left(SE(\tilde{\beta}_i)/SE(\hat{\beta}_i)\right)^2$, pour représenter la précision de la prédiction par rapport à la précision de l'estimation (Baker et Kramer [2014]).

Un critère de substitution candidat S est valide au niveau de l'étude comme un substitut au critère de jugement principal si les valeurs de la distance moyenne pondérée et de l'erreur type sont suffisamment faibles (proches de 0).

2.5 Étude de simulation

Pour évaluer les caractéristiques opérationnelles des mesures de distance proposées, nous avons effectué une étude de simulation sous différents scénarios. La méthode de validation croisée d'un contre tous (LOOCV) et de génération de données sont implémentées dans le package `surrosurv` (Rotolo et al. [2018]) du logiciel R.

2.5.1 Génération des données

Les données individuelles du critère de substitution S et du critère de jugement principal T ont été générées à partir d'une distribution de Weibull avec un paramètre de forme égale à 1 (une distribution exponentielle). Les risques de base ont été fixés, respectivement, à $4/ln(2)$ et $8/ln(2)$ pour obtenir des durées de survie médianes de 4 ans et 8 ans pour S et T. Les effets du traitement $(\alpha_i, \beta_i)'$ avaient des moyennes égales à $\alpha = \beta = ln(HR) = ln(0{,}75)$, une variance égale à $d_\alpha^2 = d_\beta^2 = 0{,}1$ et différentes valeur de corrélations ρ_{essai} selon les scénarios. Les données ont été générées à partir d'un modèle de risques proportionnels mixte et d'un modèle de copule de Clayton comme dans l'article de Rotolo, Paoletti, Burzykowski et al. [2017] avec une censure administrative [1] indépendante à 15 ans. Ce qui a donné une proportion de censure d'environ 40% pour S et 20% pour T dans tous les scénarios.

2.5.2 Choix des scénarios

Différents scénarios ont été considérés, faisant varier les paramètres suivants : la force de l'association au niveau de l'étude entre les effets du traite-

1. La censure administrative correspond aux 'exclus vivants' c'est-à-dire les patients qui n'ont pas eu l'événement d'intérêt pendant la période d'observation.

ment sur les deux critères avec le coefficient de détermination R^2, le nombre d'essais N, le nombre de patients par essai n_i, $(i = 1, \ldots, N)$, et la force de l'association au niveau individuel entre les deux critères avec le τ de Kendall.
Chaque jeu de données généré représente une méta-analyse avec un nombre d'essais randomisés $N = 10$, 20, ou 40. La taille moyenne des essais a été $n = 400$, 200, ou 100, en effet, la taille réelle de chaque essai n_i a été générée de manière aléatoire à partir d'une distribution uniforme entre $0,5 \times n$ et $1,5 \times n$. Ces différents choix de paramètres permettent d'étudier l'effet possible du nombre d'essais inclus dans une méta-analyse et de leur taille sur les mesures de distance proposées. Selon la force d'association au niveau individuel et au niveau de l'étude, quatre scénarios ont été considérés prenant en compte deux valeurs extrêmes d'association : une valeur modérée $\tau = 0,4$ et une valeur élevée $\tau = 0,6$ au niveau individuel, une valeur faible $R^2 = 0,2$ et une valeur élevée $R^2 = 0,8$ au niveau de l'étude. Ainsi, douze configurations possibles présentées dans le tableau 2.1 suivant :

TABLEAU 2.1 : *Scénarios de l'étude de simulation*

	Scénario 1	Scénario 2	Scénario 3	Scénario 4
R^2	0,2	0,2	0,8	0,8
τ de Kendall	0,4	0,6	0,4	0,6
	10 (400)	10 (400)	10 (400)	10 (400)
N (n)	20 (200)	20 (200)	20 (200)	20 (200)
	40 (100)	40 (100)	40 (100)	40 (100)

N : nombre d'essais, n : taille des essais

Pour chaque scénario, 500 réplications ont été effectuées. Pour chaque réplication, les distances moyennes pondérées ont été calculées en utilisant le modèle en deux étapes et la copule de Clayton avec et sans ajustement par les

erreurs d'estimation dans la deuxième étape. Pareillement, l'erreur type moyenne (ETM) et l'erreur type empirique (ETE) ont été rapportées sur les 500 distances obtenues pour chaque scénario.

2.5.3 Résultats

Les résultats des simulations utilisant les données générées avec le modèle de Clayton ajusté sont présentés dans cette section (tableau 2.2 et figure 2.3). Les résultats des simulations avec le modèle non ajusté sont disponibles en Annexe A (tableau A.2 et figure A.1).

Le tableau 2.2 résume la moyenne, par scénarios, des distances moyennes pondérées, leurs erreurs types moyennes et empiriques sur les 500 réplications. La figure 2.3 illustre la distribution empirique des distances moyennes pondérées en fonction de l'association au niveau de l'étude (le coefficient de détermination R^2) et l'association au niveau individuel (le τ de Kendall).

La comparaison des scénarios 1 $versus$ 3 et 2 $versus$ 4, avec une même valeur du τ de Kendall et des valeurs différentes du R^2 (0,2 et 0,8), montre que les distances moyennes pondérées étaient plus petites et les erreurs types étaient plus faibles pour les scénarios 3 et 4 avec $R^2 = 0,8$. Cela signifie que plus l'association au niveau de l'étude est élevée plus la prédiction de l'effet du traitement est précise. D'autre part, en comparant les scénarios 1 $versus$ 2 et 3 $versus$ 4, avec une même valeur du R^2 et des valeurs différentes du τ de Kendall (0,4 et 0,6), nous remarquons que les résultats sont similaires entre ces scénarios. Ainsi, les distances moyennes étaient largement indépendantes de l'association au niveau individuel. En outre, la comparaison des trois lignes de la figure 2.3 ($N = 10$; $n = 400$), ($N = 20$; $n = 200$), et ($N = 40$; $n = 100$) suggère que le nombre et la taille des essais n'ont également pas d'impact sur les mesures de distances moyennes pondérées.

Pour une valeur d'association au niveau de l'étude élevée $R^2 = 0,8$ (scénarios 3 et 4), les distances moyennes pondérées étaient en moyenne égales à 0,2. Pour une faible association au niveau de l'étude $R^2 = 0,2$ (scénarios 1 et 3), les distances étaient en moyenne égales à 0,3. A titre illustratif, ces valeurs signifient que si l'estimation de l'effet du traitement sur la survie globale $\widehat{HR}_{SG}$ est de 0,75 et l'erreur de prédiction (la distance moyenne pondérée) est de +/−0,2, la prédiction moyenne de l'effet du traitement sur la survie globale est $\widetilde{HR}_{SG} = 0,75 \times \exp(-0,2) = 0,61$ ou $\widetilde{HR}_{SG} = 0,75 \times \exp(0,2) = 0,92$ (équivalent à un $\tilde{\beta}_{SG} = ln(0,75) - 0,2 = ln(0,61)$ ou $\tilde{\beta}_{SG} = ln(0,75) + 0,2 = ln(0,92)$). De même, si l'estimation de l'effet du traitement $\widehat{HR}_{SG}$ est de 0,75 et que l'erreur de prédiction est de +/−0,3 alors la prédiction moyenne $\widetilde{HR}_{SG}$ est de 0,56 ou 1,01. La distance absolue moyenne pondérée par la somme des écarts types $\overline{diff}\left(1/\left(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i)\right)\right)$ montre la plus grande différence de valeur avec la variation de la force d'association au niveau de l'étude entre les scénarios 2 ($\tau = 0,6$; $R^2 = 0,2$) et 4 ($\tau = 0,6$; $R^2 = 0,8$). En particulier, avec un nombre d'essais de $N = 10$ de taille $n = 400$, la valeur de $\overline{diff}\left(1/\left(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i)\right)\right)$ est égale à 0,306 dans le scénario 2 et 0,169 dans le scénario 4, ce qui nous donne une différence de 0,137. Néanmoins, de façon générale, les résultats ne sont pas très différents selon l'approche de pondération.
Dans la plupart des scénarios, l'ETM est deux fois plus élevée que l'ETE, ce qui indique que les erreurs types de ces distances moyennes pondérées sont surestimées et que les intervalles de confiances estimés sont très larges.

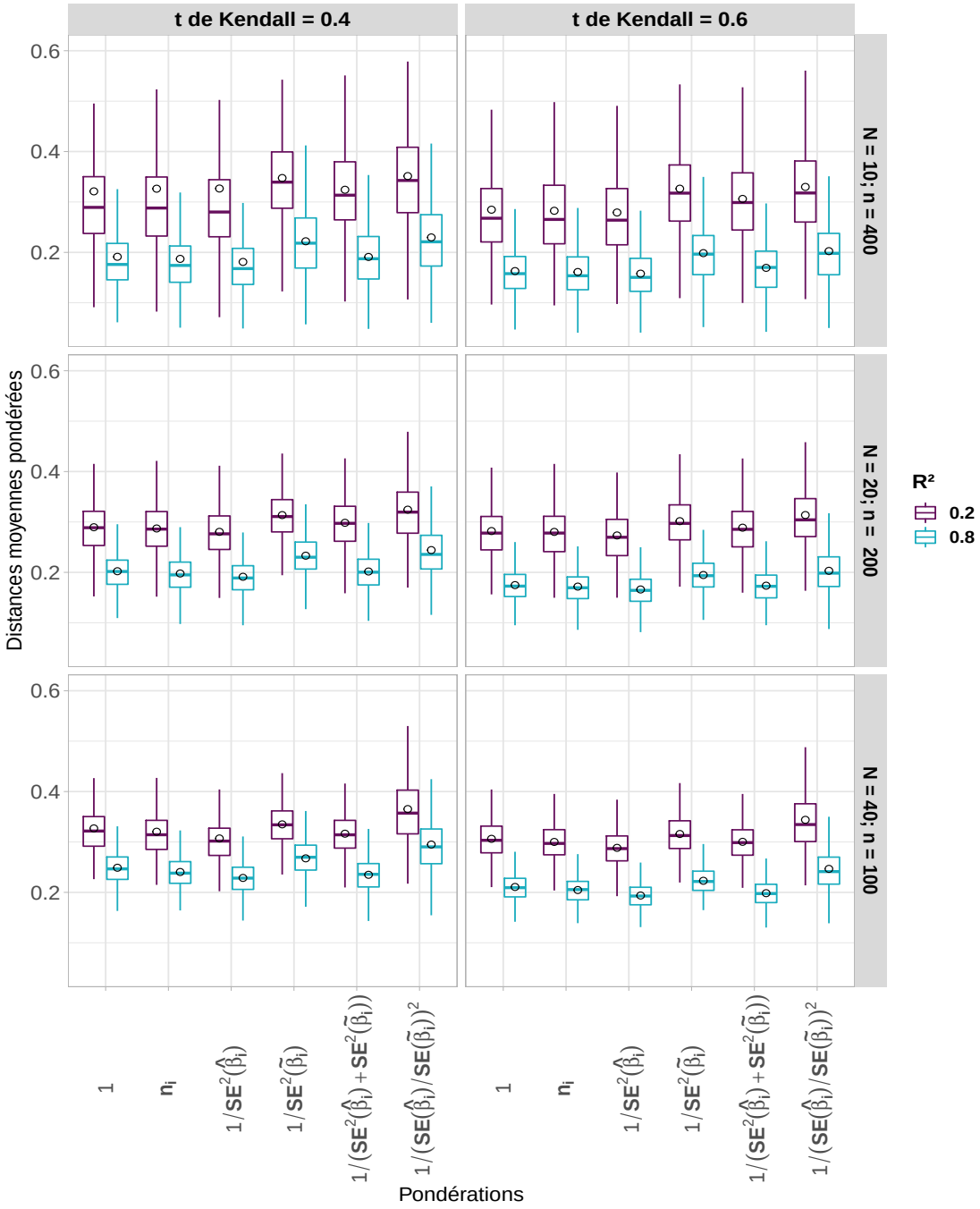
Concernant les résultats obtenus en utilisant la copule de Clayton non ajustée (Annexe, tableau A.2 et figure A.1), l'interprétation est similaire même si des distances moyennes et des erreurs types sont globalement plus faibles que pour les résultats avec la copule de Clayton ajustée.

TABLEAU 2.2 : *Distances moyennes pondérées ($\overline{diff}(w)$), erreurs types empiriques (ETE) et erreurs types moyennes (ETM) avec les données générées avec la copule de Clayton ajustée*

		scénario 1 ($\tau = 0,4; R^2 = 0,2$)			**scénario 2** ($\tau = 0,6; R^2 = 0,2$)			**scénario 3** ($\tau = 0,4; R^2 = 0,8$)			**scénario 4** ($\tau = 0,6; R^2 = 0,8$)		
	w_i	$\overline{diff}(w)$	ETE	(ETM)	$\overline{diff}(w)$	ETE	(ETM)	$\overline{diff}(w)$	ETE	(ETM)	$\overline{diff}(w)$	ETE	(ETM)
	1	0,321	0,373	(0,302)	0,285	0,102	(0,220)	0,191	0,125	(0,163)	0,163	0,054	(0,125)
	n_i	0,326	0,481	(0,291)	0,283	0,101	(0,206)	0,187	0,123	(0,151)	0,161	0,054	(0,118)
$N = 10$;	$1/SE^2(\hat{\beta}_i)$	0,327	0,535	(0,290)	0,279	0,104	(0,204)	0,181	0,093	(0,145)	0,158	0,056	(0,115)
$n = 400$	$1/SE^2(\tilde{\beta}_i)$	0,347	0,093	(0,218)	0,326	0,095	(0,206)	0,222	0,077	(0,132)	0,198	0,069	(0,118)
	$1/(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i))$	0,324	0,087	(0,214)	0,306	0,089	(0,202)	0,191	0,061	(0,129)	0,169	0,051	(0,115)
	$1/(SE(\tilde{\beta}_i)/SE(\hat{\beta}_i))^2$	0,351	0,107	(0,217)	0,330	0,103	(0,204)	0,230	0,085	(0,134)	0,202	0,074	(0,118)
	1	0,290	0,052	(0,216)	0,282	0,056	(0,215)	0,202	0,038	(0,157)	0,175	0,034	(0,134)
	n_i	0,287	0,053	(0,208)	0,280	0,061	(0,208)	0,198	0,038	(0,149)	0,172	0,034	(0,128)
$N = 20$;	$1/SE^2(\hat{\beta}_i)$	0,280	0,052	(0,203)	0,273	0,065	(0,203)	0,191	0,037	(0,144)	0,166	0,032	(0,123)
$n = 200$	$1/SE^2(\tilde{\beta}_i)$	0,314	0,051	(0,223)	0,302	0,053	(0,217)	0,233	0,049	(0,163)	0,195	0,038	(0,138)
	$1/(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i))$	0,298	0,052	(0,214)	0,289	0,053	(0,209)	0,202	0,039	(0,149)	0,174	0,034	(0,128)
	$1/(SE(\tilde{\beta}_i)/SE(\hat{\beta}_i))^2$	0,324	0,067	(0,226)	0,314	0,069	(0,223)	0,244	0,060	(0,170)	0,203	0,047	(0,143)
	1	0,327	0,119	(0,268)	0,306	0,039	(0,238)	0,249	0,033	(0,199)	0,211	0,027	(0,168)
	n_i	0,320	0,117	(0,259)	0,300	0,039	(0,229)	0,240	0,033	(0,189)	0,205	0,027	(0,160)
$N = 40$;	$1/SE^2(\hat{\beta}_i)$	0,307	0,113	(0,247)	0,288	0,037	(0,218)	0,229	0,031	(0,178)	0,194	0,025	(0,149)
$n = 100$	$1/SE^2(\tilde{\beta}_i)$	0,335	0,040	(0,254)	0,316	0,039	(0,240)	0,267	0,041	(0,202)	0,223	0,029	(0,170)
	$1/(SE^2(\hat{\beta}_i) + SE^2(\tilde{\beta}_i))$	0,316	0,040	(0,240)	0,300	0,039	(0,227)	0,235	0,033	(0,182)	0,198	0,026	(0,152)
	$1/(SE(\tilde{\beta}_i)/SE(\hat{\beta}_i))^2$	0,365	0,066	(0,276)	0,344	0,065	(0,261)	0,295	0,059	(0,223)	0,247	0,048	(0,190)

N : nombre d'essais, n : taille moyenne des essais, $\hat{\beta}$: effet observé du traitement sur le critère de jugement principal, $\tilde{\beta}$: effet prédit du traitement sur le critère de jugement principal, SE : erreur type. Quantités moyennes basées sur 500 réplications

FIGURE 2.3 : *Boxplots de distances moyennes pondérées selon les différents scénarios avec les données générées avec la copule de Clayton ajustée*



La boîte est délimitée par les 25e et 75e percentiles, contient 50% des observations et est divisée par la médiane (trait horizontal à l'intérieur de la boîte). Les moustaches de style Tukey (traits fins horizontaux) s'étendent jusqu'à un maximum de 1,5 *de l'intervalle interquartile au-delà de la boîte. Les cercles noirs sont les moyennes des distances sur 500 réplications. N : nombre d'essais, n : taille moyenne des essais, $\hat{\beta}$: effet observé du traitement sur le critère de jugement principal, $\tilde{\beta}$: effet prédit du traitement sur le critère de jugement principal, SE : erreur type*

2.6 Application : méta-analyse GASTRIC

Les différentes distances moyennes pondérées proposées dans ce premier axe de recherche ont été appliquées à des données réelles. Il s'agit de la méta-analyse GASTRIC (présentée dans la section 2.3) évaluant le bénéfice potentiel des différents schémas de chimiothérapies sur la survie globale et la survie sans progression.

TABLEAU 2.3 : *Distances moyennes pondérées entre l'effet observé et l'effet prédit du traitement sur la survie globale pour la méta-analyse GASTRIC du cancer digestif avancé*

w_i	$\overline{diff}(w)$	$SE\left(\overline{diff}(w)\right)$
1	0,234	0,178
n_i	0,202	0,154
$1/SE^2(\hat{\beta}_i)$	0,201	0,154
$1/SE^2(\tilde{\beta}_i)$	0,213	0,159
$1/\left(SE^2(\hat{\beta}_i)+SE^2(\tilde{\beta}_i)\right)$	0,201	0,155
$1/\left(SE(\tilde{\beta}_i)/SE(\hat{\beta}_i)\right)^2$	0,251	0,173

$\hat{\beta}$: effet observé du traitement sur la survie globale, $\tilde{\beta}$: effet prédit du traitement sur la survie globale, SE : erreur type

Les résultats de l'analyse présentés dans le tableau 2.3 montrent que les distances moyennes pondérées sont toutes de l'ordre de 0,2. Bien que ces distances ne soient pas censées être un moyen de recalculer le coefficient de détermination R^2, la comparaison de ces résultats avec un scénario de simulation similaire ($\tau = 0,6$, $N = 20$ et $n = 200$) montre que les distances moyennes se situent entre le scénario 4 avec $R^2 = 0,2$ et le scénario 6 avec $R^2 = 0,8$. Cela suggère une association modérée au niveau de l'étude dans les données de la

méta-analyse GASTRIC du cancer digestif avancé. Cette suggestion est cohérente avec l'estimation de $R^2 = 0,6$ publiée dans l'article de Rotolo, Paoletti, Burzykowski et al. [2017] mais qui était accompagnée d'un intervalle de confiance couvrant presque tout le support [0, 1].

2.7 Conclusion

L'approche méta-analytique de validation d'un critère de substitution est destinée à estimer un coefficient de détermination R^2 au niveau de l'étude. Néanmoins, elle fournit également un modèle de prédiction de l'effet du traitement sur le critère de jugement principal des nouveaux essais pour lesquels l'effet du traitement sur le critère de substitution a été estimé. Avec la méthode de validation croisée d'un contre tous (LOOCV), les effets estimés et les effets prédits indépendamment peuvent être comparés pour chaque essai.

Les articles de Michiels et al. [2009] et Rotolo, Pignon et al. [2017] ont rapporté le taux d'essais dont l'effet observé du traitement se situe dans l'intervalle de prédiction comme mesure de la qualité de la prédiction. On peut noter que le simple fait que la valeur de l'effet observée se situe ou non dans l'intervalle de prédiction est important, mais il ne résume pas toutes les informations concernant la cohérence entre la prédiction et l'estimation finale. Comme mentionné précédemment, le modèle en deux étapes avec la copule de Clayton ajustée ne converge pas souvent dans de nombreuses applications et, lorsqu'il converge, les intervalles de confiance de l'estimation R^2 sont souvent extrêmement larges. Ce qui nous a incités à proposer des mesures alternatives au coefficient de détermination R^2 basées sur l'erreur de prédiction absolue.

Les mesures de distance présentées dans ce travail de thèse permettent de mesurer la force de la corrélation entre les critères de jugement de type survie

en termes d'erreur de prédiction, c'est-à-dire la différence moyenne entre l'effet observé et prédit du traitement sur le critère de jugement principal.

Les résultats des simulations montrent que les distances proposées sont indépendantes de l'association au niveau individuel (τ de Kendall) quel que soit le scénario étudié (nombre d'essais $N = 10, 20, 40$ et taille des essais $n = 400, 200, 100$), ce qui est une propriété intéressante. Les distances moyennes pondérées sont plus faibles et plus précises pour une association élevée au niveau de l'étude ($R^2 = 0,8$) qu'elles ne le sont pour une faible association ($R^2 = 0,2$). Dans l'ensemble, il n'y a pas beaucoup de différence entre les résultats des modèles de copules ajustés et non ajustés.

La pondération des distances avec $1/SE(\hat{\beta})$ donne plus d'importance aux grands essais qui ont des estimations plus précises du β. La pondération avec $1/SE(\tilde{\beta})$ donne plus de poids aux essais avec une prédiction plus précise, ce qui peut signifier que ces essais ont un effet du traitement sur le critère de substitution plus proche de la moyenne de l'ensemble des essais, mais en revanche, peut également signifier plus de poids sur les petits essais pour lesquels le modèle de prédiction est estimé en excluant moins de patients. Sur la base des résultats de l'étude de simulation, il n'y a pas de différence évidente entre les stratégies de pondération, mais l'inverse de la somme des erreurs types $\overline{diff}\left(1/\left(SE^2(\hat{\beta}) + SE^2(\tilde{\beta})\right)\right)$ semble être la distance la plus discriminante entre une association faible et élevée au niveau de l'étude.

Notre étude de simulation a quelques limites. Nous avons étudié deux scénarios principaux dans l'étude de simulation : un petit nombre de grands essais et un grand nombre de petits essais. D'autres scénarios peuvent être étudiés avec un petit nombre (≤ 10) de petits essais (≤ 100). Néanmoins, ces scénarios seraient très peu informatifs et ne pourraient guère être utiles pour valider un critère de substitution. Une méta-analyse avec un grand nombre (≥ 40) de

grands essais (≥ 400) nécessitera des calculs intensifs En effet, plus le nombre d'essais dans une méta-analyse augmente plus le processus de la validation croisée est long. Par exemple, une LOOCV avec une méta-analyse de $N = 20$ essais chacun de taille $n = 200$ nécessite environ 35 minutes de calcul parallèle sur 8 coeurs. D'autre part, nous n'avons pas fait varier la variance des effets du traitement car nous ne nous attendions pas à un impact majeur sur les résultats.

Notez que la précision de la mesure R^2 et de la mesure d'erreur de prédiction proposée est souvent faible. Ce problème reflète en grande partie la rareté inhérente des informations due au nombre limité d'essais qui sont généralement disponibles dans la pratique. Par ailleurs, nous avons illustré la corrélation entre l'incertitude d'estimation (erreur type empirique (ETE)) du R^2 et les distances proposées avec le modèle de copule ajusté dans tous les scénarios (voir les figures A.2, A.3, et A.4 de l'annexe A). Les résultats ne montrent pas une corrélation clairement établie entre les deux mesures d'incertitude.

La taille d'un nouvel essai, qu'elle soit plus grande ou plus petite par rapport aux tailles des essais utilisés pour estimer le modèle, peut avoir un impact sur l'ETE de l'effet prédit du traitement sur le critère de substitution α, et donc sur la précision de la prédiction. Ce problème potentiel est commun aux mesures de validation d'un critère de substitution telles que l'effet seuil de substitution (*surrogate threshold effect (STE)*) défini comme "l'effet minimal du traitement sur le critère de substitution nécessaire pour prédire un effet non nul sur le critère de jugement principal" (Burzykowski et Buyse [2006]). Le STE dépend fortement de la variance de la prédiction ainsi des caractéristiques de la méta-analyse et du nouvel essai (Mauguen et al. [2013]).

Bien que les résultats soient cohérents dans tous les scénarios, il est difficile de recommander un seuil pour la décision finale, tout comme pour le co-

efficient R^2. Néanmoins, les mesures alternatives proposées sont plus stables et plus facilement interprétables pour un clinicien que le coefficient de détermination R^2. En outre, la distance moyenne pondérée fourni des informations supplémentaires importantes sur la qualité de prédiction et la moyenne des $\widetilde{HR}$ prédits peut être calculée et comparée à la moyenne des $\widehat{HR}$ estimés. D'après notre expérience, il est difficile pour un clinicien d'interpréter la quantité de variation expliquée pour valider un critère de substitution. Les différentes distances proposées sont mesurées sur l'échelle de l'effet du traitement. Nous proposons d'illustrer pour les cliniciens l'erreur de prédiction pour un effet du traitement estimé en termes de valeurs moyennes pour la sous-prédiction ou surprédiction des effets du traitement.

L'objectif de ce travail était de proposer une méthode alternative pour évaluer le critère de substitution de type survie pour le critère de jugement principal, basée sur l'erreur de prédiction. Bien que ce cadre soit le plus courant en oncologie, ces mesures peuvent être appliquées de manière plus générale, quel que soit le type du critère de jugement.

Ce premier axe de travail de thèse a abouti à un article publié dans la revue scientifique *Contemporary clinical trials communications* : "An alternative trial-level measure for evaluating failure-time surrogate endpoints based on prediction error".

2.8 Références

CROWLEY, J. & HOERING, A. (2012). *Handbook of statistics in clinical oncology* (3rd Edition). CRC Press. 23.

KELLY, W. & HALABI, S. (2010). *Oncology Clinical Trials : Successful Design, Conduct, and Analysis*. New York : Demos Medical Publishing. 23.

KRAMAR, A. & MATHOULIN-PÉLISSIER, S. (2011). *Methodes biostatistiques appliquées à la recherche clinique en cancérologie* (T. 5). John Libbey Eurotext. 23, 25.

BOKU, N., YAMAMOTO, S., FUKUDA, H., SHIRAO, K., DOI, T., SAWAKI, A., KOIZUMI, W., SAITO, H., YAMAGUCHI, K., TAKIUCHI, H. et al. (2009). Fluorouracil versus combination of irinotecan plus cisplatin versus S-1 in metastatic gastric cancer : a randomised phase 3 study. *The lancet oncology*, *10*(11), 1063-1069. 24.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials : definition and operational criteria. *Statistics in medicine*, *8*(4), 431-440. https://doi.org/10.1002/sim.4780080407. 25

FREEDMAN, L. S., GRAUBARD, B. I. & SCHATZKIN, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in medicine*, *11*(2), 167-178. 26.

COX, D. R. (2000). Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, *34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. 27, 60

BUYSE, M., MOLENBERGHS, G., PAOLETTI, X., OBA, K., ALONSO, A., VAN DER ELST, W. & BURZYKOWSKI, T. (2016). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal*, *58*(1), 104-132. 27.

Buyse, M. & Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics, 54*(3), 1014-29. https://doi.org/10.2307/2533853. 27

Daniels, M. J. & Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in medicine, 16*(17), 1965-1982. 28.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. & Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics, 1*(1), 49-67. https://doi.org/10.1093/biostatistics/1.1.49. 28, 30

Gail, M. H., Pfeiffer, R., Van Houwelingen, H. C. & Carroll, R. J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics, 1*(3), 231-246. 28.

Burzykowski, T., Molenberghs, G. & Buyse, M. (2006). *The evaluation of surrogate endpoints*. Springer Science & Business Media. 28.

Van Houwelingen, H. C., Arends, L. R. & Stijnen, T. (2002). Advanced methods in meta-analysis : multivariate approach and meta-regression. *Statistics in medicine, 21*(4), 589-624. https://doi.org/10.1002/sim.1040. 30

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H. & Renard, D. (2001). Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society : Series C (Applied Statistics), 50*(4), 405-422. https://doi.org/10.1111/1467-9876.00244. 30

Clayton, D. (1978). A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika, 65*(1), 141-151. https://doi.org/10.1093/biomet/65.1.141. 31

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81-93. http://www.jstor.org/stable/2332226. 32

van HOUWELINGEN, H. C., ARENDS, L. & STIJNEN, T. (2002). Advanced Methods in Meta-Analysis : Multivariate Approach and Meta-Regression. *Statistics in Medicine*, *21*, 589-624. https://doi.org/10.1002/sim.1040. 33

MICHIELS, S., LE MAÎTRE, A., BUYSE, M., BURZYKOWSKI, T., MAILLARD, E., BOGAERTS, J., VERMORKEN, J. B., BUDACH, W., PAJAK, T. F., ANG, K. K. et al. (2009). Surrogate endpoints for overall survival in locally advanced head and neck cancer : meta-analyses of individual patient data. *The lancet oncology*, *10*(4), 341-350. https://doi.org/10.1016/S1470-2045(09)70023-3. 34, 49

GROUP, G. (A. S. T. R. I. C. et al. (2013). Role of chemotherapy for advanced/-recurrent gastric cancer : an individual-patient-data meta-analysis. *European Journal of Cancer*, *49*(7), 1565-1577. https://doi.org/10.1016/j.ejca.2012.12.016. 35

PAOLETTI, X., OBA, K., BANG, Y.-J., BLEIBERG, H., BOKU, N., BOUCHÉ, O., CATALANO, P., FUSE, N., MICHIELS, S., MOEHLER, M. et al. (2013). Progression-free survival as a surrogate for overall survival in advanced/recurrent gastric cancer trials : a meta-analysis. *Journal of the National Cancer Institute*, *105*(21), 1667-1670. https://doi.org/10.1093/jnci/djt269. 35

ROTOLO, F., PAOLETTI, X., BURZYKOWSKI, T., BUYSE, M. & MICHIELS, S. (2017). A Poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses. *Statistical Methods in Medical Research*. https://doi.org/10.1093/jnci/djt269. 35, 42, 49

BURZYKOWSKI, T. & ABRAHANTES, J. C. (2005). Validation in the case of two failure-time endpoints. *The evaluation of surrogate endpoints* (p. 163-194). Springer. https://doi.org/10.1111/1467-9868.00347. 40

ROTOLO, F., PAOLETTI, X. & MICHIELS, S. (2018). Surrosurv : an R package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Computer Methods and Programs in Biomedicine, 155*, 189-198. 42.

ROTOLO, F., PIGNON, J.-P., BOURHIS, J., MARGUET, S., LECLERCQ, J., TONG NG, W., MA, J., CHAN, A. T., HUANG, P.-Y., ZHU, G. et al. (2017). Surrogate end points for overall survival in loco-regionally advanced nasopharyngeal carcinoma : an individual patient data meta-analysis. *JNCI : Journal of the National Cancer Institute, 109*(4). https://doi.org/10.1093/jnci/djw239. 49

MAUGUEN, A., PIGNON, J.-P., BURDETT, S., DOMERG, C., FISHER, D., PAULUS, R., MANDREKAR, S. J., BELANI, C. P., SHEPHERD, F. A., EISEN, T. et al. (2013). Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer : a re-analysis of meta-analyses of individual patients' data. *The lancet oncology, 14*(7), 619-626. https://doi.org/10.1016/S1470-2045(13)70158-X. 51

Chapitre 3

Prise en compte de groupes de variables dans les modèles de régression pénalisés

Sommaire

3.1 Introduction 59

3.2 Méthodes statistiques en grande dimension 60

3.2.1 Lasso standard 61

3.2.2 Sélection de variables à deux niveaux 64

3.3 Approches proposées 65

3.3.1 Lasso adaptatif 65

3.3.2 Pondérations proposées 66

3.3.2.1 Poids définis sur des coefficients de régression 67

3.3.2.1.1 Moyenne des coefficients 67

3.3.2.1.2 Analyse en composantes principales 68

3.3.2.1.3 Lasso+PCA 68

3.3.2.2 Poids définis sur la statistique de Wald 70
3.3.2.2.1 Statistique de Wald univariée 70
3.3.2.2.2 Moyenne de Wald univariée 71
3.3.2.2.3 Maximum de Wald univariée 71
3.3.2.2.4 Produit de ASW par SW 72
3.3.2.2.5 Produit de MSW par SW 73
3.4 Méthodes alternatives 73
3.4.1 Composite Minimax Concave Penalty (cMCP) 74
3.4.2 Group exponential lasso (gel) 75
3.4.3 Sparse Group Lasso (SGL) 76
3.4.4 Integrative Lasso with Penalty Factors (IPF-Lasso) . . 76
3.5 Comparaison des méthodes étudiées 77
3.5.1 Simulation de données 77
3.5.2 Choix des scénarios 79
3.5.3 Paramètres des méthodes pénalisées 80
3.5.4 Critères d'évaluation 82
3.5.5 Implémentation 84
3.5.6 Résultats 84
3.6 Application : cancer du sein précoce 96
3.7 Discussion 101
3.8 Conclusion 104
3.9 Références 105

3.1 Introduction

Le cancer du sein est le type de cancer le plus fréquent chez les femmes en France. Il représente la première cause de décès par cancer chez les femmes. Il existe diverses méthodes pour traiter le cancer du sein telles que la chimiothérapie, la chirurgie, la radiothérapie, l'hormonothérapie, etc. Bien que ces différents traitements ont amélioré le taux de survie, prédire la survie d'une patiente en fonction de ces différentes caractéristiques, tels que les biomarqueurs génomiques, restent difficiles. Pourtant, l'identification de ces biomarqueurs peut conduire à définir des profils de patientes et à améliorer leur prise en charge.
En 1987, Slamon et al. [1987] a démontré que le gène $HER2/neu$ récepteur pour les facteurs de croissance épidermiques humains (*Human Epidermal Growth Factor Receptor-2*) était amplifié de 2 à plus de 20 fois dans environ 30% des cancers du sein, et qu'il est significativement associé à la fois au délai avant rechute et à la survie globale des patientes. En effet, l'amplification du gène $HER2/neu$ a une grande valeur pronostique, les femmes atteintes d'un cancer du sein présentant une surexpression de $HER2/neu$ ont une forme agressive de la maladie avec une survie globale réduite par rapport à celles qui n'avaient aucune expression à 5 ans (34% contre 41%) (Roa et al. [2014]).
Le développement des technologies génomiques à haut débit a permis une croissance rapide et une disponibilité plus facile des données génomiques de grande dimension. Les informations obtenues pourraient éventuellement prédire l'effet du traitement (biomarqueurs prédictifs) ou le pronostic de chaque patiente (biomarqueurs pronostiques). Ces données sont collectées rapidement et à faible coût auprès des patients. Elles sont généralement caractérisées par un grand nombre de variables (p) et une petite taille d'échantillon (n) avec $p \gg n$.

Les méthodes traditionnelles de sélection de variables, telles que les méthodes de régression pas à pas (*stepwise regression*), ne sont pas adaptées à cette situation. Principalement basées sur une théorie asymptotique, elles fonctionnent bien pour un nombre de paramètres à estimer p inférieur au nombre d'observations n (un grand ratio n/p). Par conséquent, pour les données génomiques de grande dimension où $p \gg n$, se posent les problèmes de multicolinéarité, de non-identifiabilité du modèle et de sélection des variables.

3.2 Méthodes statistiques en grande dimension

Dans le cadre de données de grande dimension, la méthode de régression pénalisée appelée aussi méthode de régularisation est une meilleure alternative à la méthode du maximum de vraisemblance. La régression pénalisée consiste à introduire un terme de pénalité à la vraisemblance du modèle afin de réduire les valeurs de coefficients vers zéro. Cela permet aux variables les moins pertinentes d'avoir un coefficient proche ou égal à zéro si la réduction est suffisamment importante. Pour le modèle de risque proportionnel de Cox (Cox [2000]), avec p variables $X = (X_1, \ldots, X_p)$ et p coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$, la fonction de log-vraisemblance partielle pénalisée $\ell_p(\beta, X)$ s'écrit comme suit :

$$\ell_p(\beta, X) = \ell(\beta, X) - p(\lambda, \beta), \tag{3.1}$$

où la fonction $\ell(\beta, X)$ est la log-vraisemblance partielle, la fonction $p(\lambda, \beta)$ représente le terme de pénalité qui contrôle la complexité du modèle et λ est le paramètre de pénalisation appelé aussi paramètre de régularisation (*shrinkage*). Plus la pénalisation λ est faible plus le modèle est complexe avec un faible biais et une forte variance (surapprentissage). Inversement, plus la pé-

nalisation λ est élevée, plus le modèle est simple avec une faible variance et un fort biais (sous-apprentissage). Par conséquent, la maximisation de la log-vraisemblance partielle pénalisée 3.1 revient à chercher le meilleur compromis (biais-variance) entre la qualité de l'ajustement et la complexité du modèle.

3.2.1 Lasso standard

La méthode lasso (*Least Absolute Shrinkage and Selection Operator*) introduite par Robert Tibshirani en 1996 (TIBSHIRANI [1996]) est une méthode populaire de régression pénalisée. Le terme de pénalité du lasso correspond à la norme L_1 des coefficients de régression, et est défini par :

$$p(\lambda, \beta) = \lambda \|\beta\|_1 = \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.2}$$

Ainsi, la fonction de log-vraisemblance partielle pénalisée de la méthode lasso standard s'écrit comme :

$$\ell_p(\beta, X) = \ell(\beta, X) - \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.3}$$

La méthode lasso standard est très largement utilisée dans le cadre de données de grande dimension. La pénalité L_1 permet de sélectionner les variables : pour des valeurs élevées du paramètre de régularisation λ, certains coefficients de régression β sont estimés égaux à zéro. Si λ tend vers l'infini, toutes les variables seront exclues du modèle et si $\lambda = 0$ l'estimateur lasso coïncidera avec l'estimateur du maximum de log-vraisemblance partielle du modèle de Cox.
Habituellement, dans le modèle de Cox, le paramètre de régularisation λ est estimé en maximisant la fonction de log-vraisemblance cross-validée (*cross-*

validated loglikelihood ou *cvl*) (VERWEIJ et HOUWELINGEN [1993]). La validation croisée (comme définie dans le chapitre 2) est une méthode statistique de validation externe d'un modèle sur de nouvelles données. Elle permet d'évaluer la performance d'un modèle tout en minimisant l'erreur d'apprentissage ou la surévaluation du modèle. Cette méthode consiste à séparer les données en un sous-échantillon d'entraînement (apprentissage) et un sous-échantillon de validation (test). Le modèle est initialement ajusté sur l'échantillon d'apprentissage puis évalué sur l'échantillon externe de validation. La variante la plus courante de la validation croisée est la validation croisée à K blocs (sous-échantillons) (*K-fold cross-validation*). La méthode *K-fold cvl* suit les étapes suivantes :

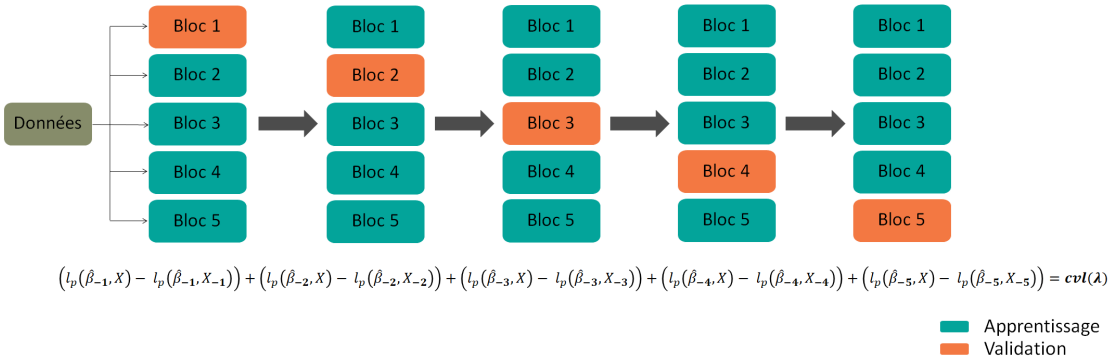
1. Diviser l'échantillon original en K blocs de taille comparable.
2. Utiliser *K-1* blocs comme sous-échantillons d'apprentissage et le bloc restant comme sous-échantillon de validation.
3. Estimer les coefficients $\widehat{\boldsymbol{\beta}}_{-k}$ en maximisant la log-vraisemblance partielle pénalisée ℓ_p sur les *K-1* blocs d'entraînement (X_{-k}). Ensuite, utiliser ces coefficients sur le K^{e} bloc externe de validation pour évaluer la contribution de sa log-vraisemblance partielle pénalisée $\ell_p\left(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_k\right)$ sur ℓ_p de l'échantillon complet $\boldsymbol{X}$. Cette contribution est calculée à partir de la différence entre les ℓ_p de l'échantillon complet et le sous-échantillon d'entraînement :
$$\ell_p\left(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}\right) - \ell_p\left(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_{-k}\right).$$
4. Répéter les étapes 2 et 3 de manière à ce que chaque sous-échantillon soit considéré comme un sous-échantillon de validation.

Après K itérations, la log-vraisemblance cross-validée *cvl* est définie comme la somme des contributions des K sous-échantillons de validation :

$$\text{cvl}(\lambda) = \sum_{k=1}^{K} \left(\ell_p \left(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X} \right) - \ell_p \left(\widehat{\boldsymbol{\beta}}_{-k}, \boldsymbol{X}_{-k} \right) \right). \tag{3.4}$$

Le paramètre de régularisation optimal $\hat{\lambda}_{cvl}$ est celui qui maximise la fonction *cvl*. La figure 3.1 montre un exemple d'itérations pour une validation croisée à K = 5 blocs. À noter qu'en pratique, selon la taille de l'échantillon et le découpage des blocs, le $\hat{\lambda}_{cvl}$ peut varier et cela peut impacter l'estimation des paramètres.

FIGURE 3.1 : *Exemple de validation croisée à 5 blocs*



Bien que la méthode lasso standard ait une bonne capacité de sélection de variables, elle possède quelques limites. Meinshausen et Bühlmann [2006], Zhao et Yu [2006] ont montré que cette méthode conduit souvent à la sélection d'un modèle avec un très grand nombre de faux positifs (FP). Des modifications particulières peuvent être utilisées comme un choix plus conservateur du λ optimal (Ternès et al. [2016]) pour pallier à ce problème. D'autre part, le lasso standard se focalise généralement sur des données homogènes en termes de nature et de technique d'acquisition, et sans aucune connaissance présumée de leurs rôles biologiques. Cependant, dans les applications réelles, il est de plus en plus courant de traiter de multiples sources de données telles que les données de séquençage, les variations du nombre de copies, les mutations ou encore les données génomiques pour lesquelles l'appartenance à différentes

voies biologiques est une information importante à prendre en compte.

3.2.2 Sélection de variables à deux niveaux

Supposons que nous avons deux groupes de biomarqueurs, l'un est **actif** (c'est-à-dire il contient certains biomarqueurs pronostiques) et l'autre groupe est **inactif** (c'est-à-dire il ne contient aucun biomarqueur associé au critère de jugement principal). Compte tenu de ces informations, il est plus judicieux de favoriser la sélection des biomarqueurs appartenant au groupe actif.
Par exemple, nos données réelles sur l'expression génique sont composées de 614 patientes atteintes d'un cancer du sein et traitées par une chimiothérapie adjuvante à base d'anthracyclines, et de 128 gènes appartenant à quatre voies biologiques [1] importantes. L'objectif principal est de prendre en compte ces informations biologiques disponibles (la structure en groupe connue a priori) dans la procédure de sélection des biomarqueurs pronostiques.
L'idée est de faire la sélection à deux niveaux : à la fois des groupes pertinents et des biomarqueurs au sein des groupes sélectionnés. Pour ce faire, nous cherchons à pénaliser les biomarqueurs qui, en plus d'avoir un faible effet, appartiennent à une voie biologique ayant un faible effet global ; et à favoriser la sélection des biomarqueurs qui, en plus d'avoir un effet important, appartiennent à une voie ayant un effet global important.
Depuis sa publication (Tibshirani [2011]), plusieurs variantes de la méthode lasso standard ont été proposées. Elles diffèrent par le choix de la fonction de pénalité. Des exemples de ces méthodes seront détaillées dans la section 3.3. Nous nous concentrons ici sur la sélection des biomarqueurs appartenant à des groupes disjoints prédéfinis.

1. Une voie biologique correspond à un ensemble de gènes qui interagissent ensemble pour assurer une fonction biologique ou un processus cellulaire.

3.3 Approches proposées

Dans ce travail de recherche, nous avons proposé différentes stratégies de pondérations destinées à la méthode lasso adaptatif pour la sélection de variables à deux niveaux. Nous présentons dans la section 3.3.1 la pénalisation lasso adaptatif et dans la section 3.3.2, les différentes méthodes d'estimations des pondérations proposées.

3.3.1 Lasso adaptatif

La méthode lasso adaptatif, proposée par Zou [2006] et adaptée par H. H. Zhang et Lu [2007] pour les données de survie, est une extension du lasso standard. La particularité de cette extension est qu'elle attribue des poids adaptatifs W pour pénaliser différemment les coefficients de régressions β de la pénalité L_1. Alors que pour le lasso standard la pénalisation est la même avec un poids $W = 1$ pour tous les biomarqueurs, la pondération lasso adaptatif s'écrit :

$$p(\lambda, \beta) = \lambda \sum_{j=1}^{p} W_j \left|\beta_j\right|. \tag{3.5}$$

La fonction de log-vraisemblance pénalisée du lasso adaptatif s'écrit alors comme suit :

$$\ell_p(\beta, X) = \ell(\beta, X) - \lambda \sum_{j=1}^{p} W_j \left|\beta_j\right|, \tag{3.6}$$

où $W = (W_1, \ldots, W_p)^T$ est le vecteur de poids positifs spécifiques aux biomarqueurs dont la valeur est choisie d'une manière adaptative, le paramètre λ est calculé par une validation croisée à K blocs. Plus le poids W_j accordé à un biomarqueur j est élevé, plus la valeur de son coefficient estimée tend vers zéro

($\widehat{\beta}_j \longrightarrow 0$) ; ainsi la chance qu'il soit sélectionné dans le modèle est faible.
Autre avantage de cette méthode : contrairement au lasso standard, l'estimateur du lasso adaptatif satisfait la propriété d'oracle. En termes simples, un modèle de régression possède la propriété d'oracle s'il satisfait les conditions suivantes (Fan et Li [2001]) :

— Pour un échantillon de taille infinie, la probabilité que la procédure estime correctement les coefficients non nuls tend vers 1.
— Les estimations des coefficients non nuls sont asymptotiquement normales avec les mêmes moyennes et covariances que si le vrai modèle était connu a priori.

La procédure du lasso adaptatif se réalise en deux étapes. À la première étape, les poids adaptatifs (W_j) sont estimés selon une méthode définie en amont. Dans leur article, H. H. Zhang et Lu [2007] ont proposé un poids $W_j = \left|\widehat{\beta}_j\right|$ où le vecteur $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^T$ est obtenu en maximisant la fonction de log-vraisemblance partielle $\ell(\beta, X)$. À la deuxième étape, ces poids spécifiques sont utilisés pour ajuster la pénalité attribuée à chacun des coefficients de régression.
Nous avons proposé plusieurs choix de pondération afin de pénaliser les biomarqueurs selon leur appartenance aux groupes actifs ou inactifs et selon leur force d'association avec le critère de jugement principal. L'objectif était d'attribuer une faible pondération (W_j) aux biomarqueurs appartenant à des groupes qui semblaient jouer un rôle pronostique (actifs), afin d'augmenter les chances qu'ils soient sélectionnés dans le modèle final.

3.3.2 Pondérations proposées

Soit R un nombre prédéfini de groupes de biomarqueurs et p_r le nombre de biomarqueurs dans le groupe r, ($r = 1, \ldots, R$), avec $p = \sum_r p_r$. La fonction

log-vraisemblance partielle de la méthode lasso adaptatif 3.6 s'écrit alors :

$$\ell_p(\beta, X) = \ell(\beta, X) - \lambda \sum_{r=1}^{R} \left(\sum_{k=1}^{p_r} W_k{}^{(r)} \left| \beta_k^{(r)} \right| \right). \tag{3.7}$$

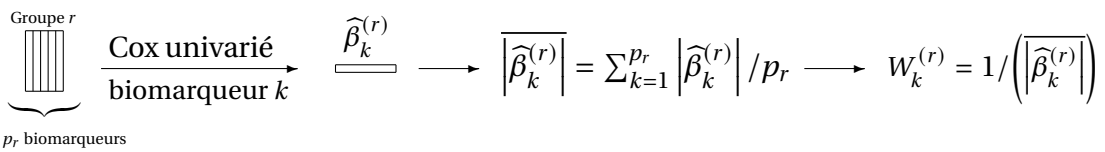
Nous avons proposé huit méthodes d'estimation des poids adaptatifs W basées sur le coefficient de régression $\widehat{\beta}$ et la statistique de test de Wald obtenue par un modèle de Cox univarié (voir ci-dessous). Chaque méthode de pondération est présentée avec un schéma explicatif.

3.3.2.1 Poids définis sur des coefficients de régression

3.3.2.1.1 Moyenne des coefficients

La moyenne des coefficients de régression (*Average Coefficients* ou AC) consiste à résumer les informations de chaque groupe de biomarqueurs en utilisant la moyenne des valeurs absolues des coefficients de régression univariée de Cox. Pour chaque biomarqueur k appartenant à un groupe donné r, un modèle de Cox univarié est ajusté pour estimer le coefficient de régression $\widehat{\beta}_k^{(r)}$. Ainsi, le poids adaptatif pour chaque biomarqueur k d'un groupe r, $(k = 1, \ldots, p_r)$, est défini par :

$$W_k^{(r)} = 1 / \left(\overline{\left| \widehat{\beta}_k^{(r)} \right|} \right) = 1 / \left(\sum_{k=1}^{p_r} \left| \widehat{\beta}_k^{(r)} \right| / p_r \right), \quad r = 1, \ldots, R. \tag{3.8}$$



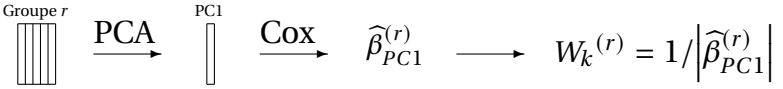
$$\left| \widehat{\beta}_k^{(r)} \right| \nearrow\nearrow \longrightarrow W_k^{(r)} \searrow\searrow$$

Notez que nous utilisons la valeur absolue de $\widehat{\beta}_k$ pour tenir compte des effets positifs et négatifs des biomarqueurs.

3.3.2.1.2 Analyse en composantes principales

L'analyse en composantes principales (*Principal Component Analysis* ou *PCA*) est une approche standard pour résumer les informations pertinentes d'un groupe de variables. La première composante principale $PC1$ est la combinaison linéaire d'un groupe de biomarqueurs avec la variance la plus élevée. La $PC1^{(r)}$, $(r = 1, \ldots, R)$ est calculée pour chaque groupe, puis le coefficient de régression $\widehat{\beta}^{(r)}_{PC1}$ est estimé avec un modèle de Cox univarié pour chaque $PC1^{(r)}$. Pour chaque biomarqueur k appartenant à un groupe donné r, le poids adaptatif se calcule ainsi :

$$W_k{}^{(r)} = 1/\left|\widehat{\beta}^{(r)}_{PC1}\right|, \quad k = 1, \ldots p_r, \quad r = 1, \ldots, R. \tag{3.9}$$



3.3.2.1.3 Lasso+PCA

Afin de sélectionner les biomarqueurs les plus pertinents du point de vue pronostique, deux étapes ont été envisagées dans cette approche :

— Dans un premier temps, la méthode lasso standard a été appliquée à tous les biomarqueurs, quelle que soit la structure en groupe, afin d'éliminer les biomarqueurs montrant la moindre association avec la survie des patients.

Soit R', $1 \leq R' \leq R$, le nombre de groupes présélectionnés, c'est-à-dire, les groupes contenant au moins un biomarqueur sélectionné lors de cette première étape. Considérons que $R - R'$ est le nombre de groupes non présélectionnés, c'est-à-dire, les groupes ne contenant aucun biomarqueur sélectionné par la méthode lasso standard lors de cette première étape.

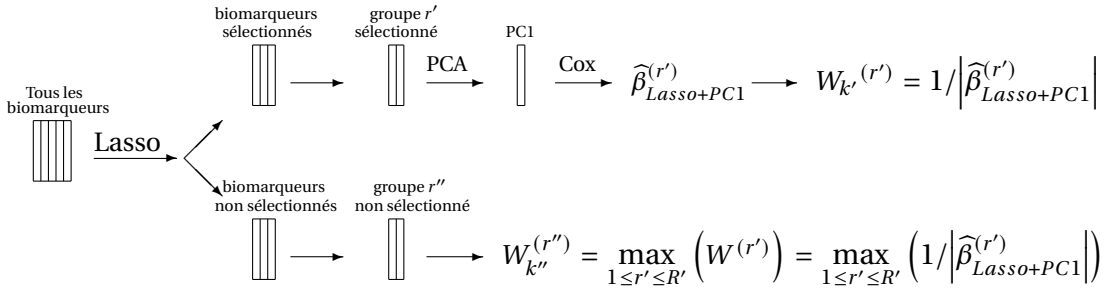
— Ensuite, la première composante principale $PC1^{(r')}$, $(r' = 1, \ldots, R')$ a été calculée pour chacun des groupes sélectionnés lors de la première étape, parmi les biomarqueurs présélectionnés lors de la première étape. Si un groupe ne contenait qu'un seul biomarqueur présélectionné, sa première composante principale serait ce biomarqueur sélectionné lui-même.

Le coefficient $\widehat{\beta}^{(r')}_{Lasso+PC1}$ est enfin estimé dans un modèle univarié de Cox n'incluant que la $PC1$ du groupe. Pour chaque biomarqueur sélectionné k' appartenant à un même groupe r', le poids adaptatif s'écrit :

$$W^{(r')}_{k'} = 1/\left|\widehat{\beta}^{(r')}_{Lasso+PC1}\right|, \quad k' = 1, \ldots p'_r, \quad r' = 1, \ldots, R'. \tag{3.10}$$

La pondération la plus élevée a été accordée aux biomarqueurs non présélectionnés. En effet, pour chaque biomarqueur non présélectionné k'' appartenant à un groupe r'', $(r'' = 1, \ldots, R - R')$, le poids adaptatif a été le même et est égal au maximum des poids accordés aux biomarqueurs présélectionnés :

$$W^{(r'')}_{k''} = \max_{1 \leq r' \leq R'} \left(W^{(r')}\right) = \max_{1 \leq r' \leq R'} \left(1/\left|\widehat{\beta}^{(r')}_{Lasso+PC1}\right|\right), \quad k' = 1, \ldots p''_r. \tag{3.11}$$

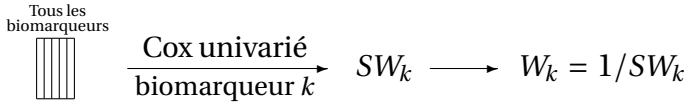


3.3.2.2 Poids définis sur la statistique de Wald

3.3.2.2.1 Statistique de Wald univariée

À titre de comparaison, la pondération statistique de Wald univariée (*Single Wald* ou *SW*) a été incluse dans ce travail de recherche. La stratégie de pondération *SW* attribue à chaque biomarqueur un poids égal à l'inverse de la valeur de sa statistique de Wald du modèle de Cox univarié.
La statistique de test de Wald appelée aussi Z-test mesure la significativité d'un coefficient de régression. Elle se fonde sur la normalité asymptotique de l'estimateur du maximum de vraisemblance partielle. Pour un coefficient de régression univariée de Cox β, l'hypothèse nulle et l'hypothèse alternative sont écrites comme $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. La statistique de Wald se calcule comme $SW = \widehat{\beta}^2/\mathbb{V}(\widehat{\beta}) \sim \chi^2(1)$ et suit une loi Chi-deux à un degré de liberté. Le poids adaptatif a été défini, pour chaque biomarqueur k, $(k = 1, \ldots, p)$ comme :

$$W_k = 1/SW_k = 1/\left(\widehat{\beta}_k^2/\mathbb{V}(\widehat{\beta}_k)\right), \quad k = 1, \ldots, p. \tag{3.12}$$

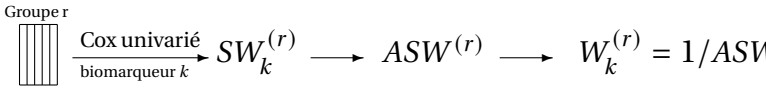


Bien que cette pondération attribue un poids différent pour chaque biomarqueur sans tenir compte de l'effet de groupe, nous avons trouvé utile d'évaluer l'effet supplémentaire de la combinaison de la pondération SW avec la moyenne de Wald univariée (*Average Single Wald* ou ASW) et le maximum de Wald univariée (*Max Single Wald* ou MSW) décrits ci-dessous.

3.3.2.2.2 Moyenne de Wald univariée

La pondération par la moyenne des statistiques de Wald univariées (*Average Single Wald* ou ASW) a été proposée pour résumer les informations d'un groupe de biomarqueurs. Dans un premier temps, pour chaque biomarqueur appartenant à un même groupe, le modèle de Cox univarié a été ajusté pour estimer la statistique de Wald (méthode ci-dessous). Ensuite, pour chaque groupe, la moyenne des statistiques de Wald univariées a été calculée $ASW^{(r)} = \sum_{k=1}^{p_r} SW_k^{(r)}/p_r$ avec $SW_k^{(r)} = \left(\widehat{\beta}_k^{(r)}\right)^2 / \mathbb{V}\left(\widehat{\beta}_k^{(r)}\right)$, $(r = 1, \ldots, R)$. Le poids adaptatif été défini par la formule suivante :

$$W_k^{(r)} = 1/ASW^{(r)} = 1/\left(\sum_{k=1}^{p_r} SW_k^{(r)}/p_r\right), \quad r = 1, \ldots, R. \tag{3.13}$$



3.3.2.2.3 Maximum de Wald univariée

Également, nous avons proposé la pondération par le maximum des statis-

tiques de Wald univariées (*Max Single Wald* ou MSW) (Michiels et al. [2011]). Pour chaque groupe de biomarqueurs, le MSW a été calculé comme $MSW^{(r)} = \max_{1\leq k\leq p_r} SW_k^{(r)}$ et le poids adaptatif s'écrit :

$$W_k^{(r)} = 1/MSW^{(r)} = 1/\left(\max_{1\leq k\leq p_r} SW_k^{(r)}\right), \quad r = 1,\ldots,R. \tag{3.14}$$

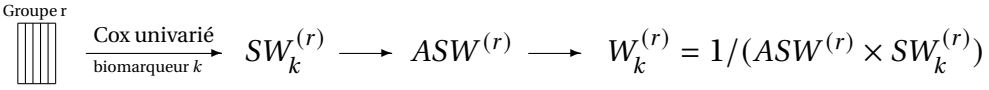
En conséquence, les groupes actifs contenant des biomarqueurs fortement associés au critère de jugement principal (avec des valeurs de $SW_k^{(r)}$ élevées ainsi que des $MSW^{(r)}$ élevées), sont très légèrement pénalisés (faibles $W_k^{(r)}$). Tandis qu'une pondération importante a été appliquée aux groupes inactifs.
À noter que les approches de pondération présentées ci-dessus attribuent des poids différents aux différents groupes, mais le même poids aux biomarqueurs du même groupe. Dans ce qui suit, nous présentons d'autres approches de pondération basées sur les statistiques de Wald qui attribuent non seulement des poids différents aux groupes de biomarqueurs mais aussi aux biomarqueurs au sein du même groupe. L'objectif de ces approches est de combiner deux poids afin de favoriser davantage la sélection à la fois des groupes et des biomarqueurs à l'intérieur des groupes sélectionnés.

3.3.2.2.4 Produit de ASW par SW

La pondération par le produit de ASW par SW (*Product of Average Single Wald by Single Wald* ou $ASW \times SW$) a été définie comme le produit de la $ASW^{(r)}$, d'un groupe r, par la $SW_k^{(r)}$, du biomarqueur k appartenant à ce groupe. Ainsi,

le poids adaptatif s'écrit :

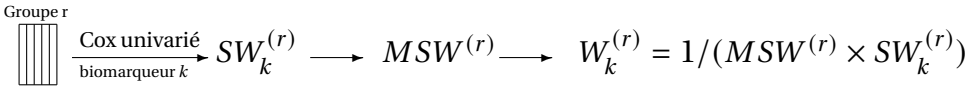
$$W_k^{(r)} = 1/(ASW^{(r)} \times SW_k^{(r)}), \quad r = 1, \ldots, R. \tag{3.15}$$



3.3.2.2.5 Produit de MSW par SW

De même, nous avons proposé la pondération (*Product of Max Single Wald by Single Wald* ou $MSW \times SW$) comme le produit du $MSW^{(r)}$ par la $SW_k^{(r)}$. De la même manière que $ASW \times SW$, le poids adaptatif a été défini sous la forme :

$$W_k^{(r)} = 1/(MSW^{(r)} \times SW_k^{(r)}), \quad r = 1, \ldots, R. \tag{3.16}$$



3.4 Méthodes alternatives

Dans cette section, nous présentons quatre méthodes alternatives à l'approche lasso adaptatif. Contrairement à cette dernière qui repose sur l'attribution de poids dans la fonction de pénalité, ces méthodes sont des variantes de la méthode lasso standard avec différentes fonctions de pénalité permettant la sélection à deux niveaux.

3.4.1 Composite Minimax Concave Penalty (cMCP)

Breheny et Huang [2009] ont proposé la pondération (*group Minimax Concave Penalty*), appelée *composite Minimax Concave Penalty* ou cMCP par Huang et al. [2012], pour effectuer la sélection de variables à deux niveaux. Le cMCP est défini comme une pénalité extérieure appliquée à la somme des pénalités intérieures, sa fonction de log-vraisemblance partielle pénalisée a la forme suivante :

$$\ell_p(\beta, X) = \ell(\beta, X) - \sum_{r=1}^{R} f_O \left\{ \sum_{k=1}^{p_r} f_I \left(\beta_k^{(r)} \right) \right\}, \tag{3.17}$$

où $f_O = f_{\lambda,b}$ et $f_I = f_{\lambda,a}$ représentent, respectivement, les fonctions de pénalité MCP (C.-H. Zhang et al. [2010]) extérieure (*outer*) et intérieure (*inner*). $a > 0$ et $b > 0$ sont, respectivement, les paramètres de forme pour les pénalités extérieure et intérieure. Simultanément, la pénalité externe réduit l'effet des groupes non importants à zéro et la fonction de pénalité interne exclut les variables non pertinentes au sein des groupes sélectionnés. L'équation montre que la pénalisation se compose de deux termes : le premier porte des informations concernant le groupe et le second porte des informations sur la variable individuelle. La sélection d'une variable dans le modèle est affectée à la fois par la force de son association individuelle avec le critère de jugement principal et par le signal collectif du groupe auquel elle appartient. Par conséquent, une variable ayant une association modérée avec le résultat peut être incluse dans le modèle si elle appartient à un groupe contenant d'autres variables ayant des fortes associations avec le résultat, ou peut être exclue si le reste des variables du même groupe présente peu d'association avec le résultat. La pénalité MCP est définie sur $[0, \infty)$ par :

$$f_{\lambda,a}(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2a} & \text{si } \beta \leq a\lambda \\ \frac{1}{2}a\lambda^2 & \text{si } \beta > a\lambda \end{cases}, \quad \text{avec } \lambda \geq 0. \tag{3.18}$$

Afin de comprendre le fonctionnement de cette pénalité, considérons sa dérivée :

$$f'_{\lambda,a}(\beta) = \begin{cases} \lambda - \frac{\beta}{a} & \text{si } \beta \leq a\lambda \\ 0 & \text{si } \beta > a\lambda. \end{cases} \tag{3.19}$$

Le MCP commence par appliquer le même taux de pénalisation que la méthode lasso standard, puis il relaxe continuellement cette pénalité jusqu'à ce que $\beta > a\lambda$, où le taux de pénalisation devient nul.

3.4.2 Group exponential lasso (gel)

La méthode gel (*group exponential lasso*) (Breheny [2015]) effectue également la sélection de variables à deux niveaux. Elle maximise une fonction de log-vraisemblance partielle pénalisée similaire à celle du cMCP, avec une pénalité extérieure égale à la pénalité exponentielle et une pénalité intérieure égale à la pénalité du lasso standard. La pénalité exponentielle est définie sur $[0, \infty)$ comme suit :

$$f_{\lambda,\tau}(\beta) = \frac{\lambda^2}{\tau}\left\{1 - \exp\left(-\frac{\tau\beta}{\lambda}\right)\right\}, \tag{3.20}$$

où τ est le taux de décroissance exponentielle. Lorsque $\tau \longrightarrow 0$, la méthode gel sélectionne un modèle équivalent au lasso standard, tandis que si $\tau \longrightarrow 1$, la

méthode gel sélectionne un modèle équivalent au lasso groupé (*group lasso*) (Yuan et Lin [2006]).

La méthode lasso groupé n'a pas été incluse dans notre étude car elle n'effectue pas la sélection des biomarqueurs individuels mais uniquement des groupes, ce qui ne répond pas à notre objectif de sélection à deux niveaux.

3.4.3 Sparse Group Lasso (SGL)

Simon et al. [2013] a proposé la méthode SGL (*Sparse Group Lasso*) pour favoriser la sélection à deux niveaux (*groupwise sparsity* et *within group sparsity*) en sélectionnant les groupes pertinents dans la même optique que la méthode lasso groupé et les biomarqueurs pertinents au sein des groupes sélectionnés. La fonction de log-vraisemblance partielle pénalisée de la méthode SGL est la suivante :

$$\begin{aligned} \ell_p(\beta, X) &= \ell(\beta, X) \; - (1-\alpha)\lambda \sum_{r=1}^{R} \sqrt{p_r} \left\| \beta^{(r)} \right\|_2 - \alpha\lambda \left\| \beta \right\|_1 \\ &= \ell(\beta, X) \; - (1-\alpha)\lambda \sum_{r=1}^{R} \sqrt{p_r \sum_{k=1}^{p_r} \beta_k^{(r)}} - \alpha\lambda \sum_{j=1}^{p} \left| \beta_j \right|, \end{aligned} \tag{3.21}$$

où le paramètre $\alpha \in [0, 1]$, en fixant $\alpha = 1$ ou $\alpha = 0$, la pénalité SGL correspond, respectivement, au lasso standard ou au lasso groupé.

3.4.4 Integrative Lasso with Penalty Factors (IPF-Lasso)

L'IPF-lasso (*Integrative Lasso with Penalty Factors*), proposée par Boulesteix et al. [2017], est une autre variante de la méthode lasso standard utilisée pour les données avec une structure en groupe. La méthode IPF-Lasso

introduit divers facteurs de pénalité pour les différents groupes afin de pondérer les données différemment selon leurs sources. La fonction de log-vraisemblance partielle pénalisée de cette méthode est définie par :

$$\ell_p(\beta, X) = \ell(\beta, X) - \sum_{r=1}^{R} \boldsymbol{\lambda}_r \left\| \beta^{(r)} \right\|_1 = \ell(\beta, X) - \sum_{r=1}^{R} \boldsymbol{\lambda}_r \sum_{k=1}^{p_r} \left| \beta_k^{(r)} \right|, \qquad (3.22)$$

où les facteurs de pénalité $\lambda_1, \lambda_2, \ldots, \lambda_R$ sont choisis par validation croisée (5 blocs et 10 répétitions) à partir d'une liste de vecteurs candidats prédéfinis. Dans nos simulations, nous avons proposé deux listes de facteurs de pénalité candidats pour tenir compte de l'effet groupes de biomarqueurs (voir ci-dessous).

3.5 Comparaison des méthodes étudiées

Pour comparer les pondérations du lasso adaptatif proposées avec la méthode lasso standard et les quatre autres compétiteurs (cMCP, gel, SGL et IPF-Lasso), une étude de simulation a été réalisée évaluant la capacité de sélection des groupes et des biomarqueurs pronostiques (actifs) et la précision de prédiction.

3.5.1 Simulation de données

Deux études de simulations ont été effectuées chacune avec un nombre de biomarqueurs (expressions de gènes) $p = 1000$ un nombre de patients $n = 500$. Pour la première étude de simulation (a), les biomarqueurs ont été répartis de manière égale sur $R = 20$ groupes, de $p_r = 50$, $(r = 1, \ldots, R)$, biomarqueurs ; Quant à la deuxième étude de simulation (b), les 20 groupes de biomarqueurs étaient de tailles différentes : 10 groupes de $p_r = 25$ biomarqueurs chacun et les

10 autres groupes de $p_r = 75$ biomarqueurs chacun. Les biomarqueurs ont été générés à partir d'une distribution normale multivariée $\boldsymbol{X} \sim N_p(\mu, \boldsymbol{\Sigma})$ centrée (les moyennes $\mu_1 = \cdots = \mu_p = 0$) et réduite (les écarts types $\sigma_1 = \cdots = \sigma_p = 1$). La matrice de variance covariance $\boldsymbol{\Sigma}$ était de taille 1000×1000.
Nous avons considéré des groupes de biomarqueurs disjoints non corrélés et prédéfinis. Pour ce faire, nous avons supposé que les biomarqueurs appartenant à des groupes différents étaient indépendants les uns des autres, mais les biomarqueurs au sein d'un même groupe étaient corrélés entre eux avec une structure de corrélation autorégressive de blocs de 5 biomarqueurs (PANG et al. [2009]). Ainsi, $\boldsymbol{\Sigma}$ était définie comme une matrice diagonale par blocs :

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_\rho & 0 & \dots & \dots & \dots & \vdots \\ 0 & \Sigma_\rho & 0 & \ddots & \ddots & \vdots \\ \vdots & 0 & \Sigma_\rho & 0 & \ddots & \vdots \\ \vdots & \ddots & 0 & \Sigma_\rho & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & \Sigma_\rho & \vdots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}_{1000\times1000} , \tag{3.23}$$

Chaque bloc de matrice Σ_ρ représente une matrice carré 5×5 :

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \dots & \rho^3 & \rho^4 \\ \rho & 1 & \ddots & \dots & \rho^3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^3 & \dots & \ddots & 1 & \rho \\ \rho^4 & \rho^3 & \dots & \rho & 1 \end{pmatrix}_{5\times5} , \tag{3.24}$$

La corrélation entre deux biomarqueurs i et j d'un même bloc a été fixée à $\rho_{ij} = 0,8^{|i-j|}$. Pour imiter les données réelles de notre application sur l'expression génique du cancer du sein, une durée de survie exponentielle a été générée

avec une survie médiane de huit ans. Le taux de censure indépendante, générée à partir d'une distribution uniforme U(2,6), variait entre 55% et 78%, reflétant un essai avec un temps de suivi de deux ans et un temps de recrutement de quatre ans. Un total de $n = 1000$ patients ont été considéré pour chaque jeu de données simulé : 500 patients ont été utilisés pour le jeu de données d'entraînement et les 500 autres pour le jeu de données de validation.
Enfin, une dernière étude de simulation a été réalisée dans un cas extrême en considérant seulement $n = 50$ patients et $p = 1000$ biomarqueurs simulés de la même manière que l'étude de simulation (a), pour évaluer la robustesse des résultats avec $n = 50$ patients par rapport aux $n = 500$ patients.

3.5.2 Choix des scénarios

Dans les deux études de simulation (a) et (b), nous avons considéré 8 scénarios avec différentes valeurs du nombre de biomarqueurs actifs q, le nombre de groupes actifs l et l'effet des biomarqueurs actifs β. Un biomarqueur actif est défini comme un biomarqueur pronostique et un groupe actif est un groupe contenant un ou plusieurs biomarqueurs actifs. Pour chaque scénario, 500 réplications ont été effectuées.
Les deux premiers scénarios 1 et 2 sont des scénarios nuls (sans aucun effet des groupes de biomarqueurs). Le scénario 1 ne considère aucun biomarqueur actif ($l = 0$ et $q = 0$). Quant au scénario 2 ($l = 20$ et $q = 100$), tous les groupes de biomarqueurs sont actifs, chaque groupe contenant le même nombre de biomarqueurs actifs, à savoir 5. L'effet des biomarqueurs actifs ou rapport de risque instantané (*Hazard Ratio*) $HR = \exp(\beta)$ a été généré aléatoirement entre 0,85 et 0,95 ($\beta \sim U(-0,05, -0,16)$).
Dans les scénarios alternatifs 3-8, les biomarqueurs actifs n'appartiennent qu'à certains groupes prédéfinis. Le nombre de biomarqueurs actifs l et de groupes

actifs q était comme suit :

— Scénario 3 : $l = 1$ et $q = 8$

— Scénario 4 : $l = 2$ et $q = 8$

— Scénario 5 : $l = 1$ et $q = 32$

— Scénario 6 : $l = 2$ et $q = 32$

— Scénario 7 : $l = 2$ et $q = 16$

— Scénario 8 : $l = 2$ et $q = 8$

Bien que les scénarios 4 et 8 aient le même nombre de biomarqueurs et de groupes actifs ($l = 2; q = 8$), ils ont des effets de biomarqueurs actifs différents : le HR a été généré de manière aléatoire entre 0, 65 et 0, 75 ($\beta \sim U(-0,43, -0,29)$) pour les effets de biomarqueurs importants dans les scénarios 3-4 et entre 0, 85 et 0, 95 ($\beta \sim U(-0,05, -0,16)$) pour des effets faibles dans les scénarios 5-8. Les biomarqueurs actifs ont été répartis de manière égale entre tous les groupes actifs. Le groupe actif et l'indice des biomarqueurs, ainsi que des informations plus détaillées sur le processus de simulation, sont présentés en Annexe B tableau B.1.

3.5.3 Paramètres des méthodes pénalisées

Les méthodes de sélection ont été appliquées à chaque ensemble de données générées. Le paramètre de régularisation λ a été sélectionné en maximisant la log-vraisemblance cross-validée (cvl) par une validation croisée à 5 blocs. Pour la méthode IPF-Lasso, ce processus de validation croisée a été répété 10 fois avec différentes partitions de 5 blocs pour avoir des résultats plus stables. Pour l'IPF-Lasso, nous avons fixé deux listes de facteurs de pénalité. La première liste (`pflist1`) attribue à un, deux ou trois groupes une pénalité deux fois inférieure à celle de tous les autres groupes, laissant la validation croisée le

choix des groupes les plus aptes à être le moins pénalisés. La deuxième liste (`pflist2`) attribuait à un, deux ou trois groupes une pénalité deux fois plus élevée que celle des autres groupes, avec là encore un choix automatique du groupe le plus apte à recevoir la pénalité la plus élevée.

```
pflist1 = list(
  c(1, rep(2, 19)),
  c(2, 1, rep(2, 18)),
  c(rep(2, 10), 1, rep(2, 9)),
  c(1, 1, rep(2, 18)),
  c(1, rep(2, 9), 1, rep(2, 9)),
  c(2, 1, rep(2, 8), 1, rep(2, 9)),
  c(1, 1, rep(2, 8), 1, rep(2, 9))
)

pflist2 = list(
  c(2, rep(1, 19)),
  c(1, 2, rep(1, 18)),
  c(rep(1, 10), 2, rep(1, 9)),
  c(2, 2, rep(1, 18)),
  c(2, rep(1, 9), 2, rep(1, 9)),
  c(1, 2, rep(1, 8), 2, rep(1, 9)),
  c(2, 2, rep(1, 8), 2, rep(1, 9))
)
```

Ayant simulé des biomarqueurs actifs dans un ou deux groupes maximum (le 1er, le 2e ou le 11e voir Annexe B tableau B.1), la configuration des listes de facteurs de pénalisation permet d'étudier les cas où un nombre insuffisant, le bon nombre, ou un nombre excessif de groupes sont pénalisés différemment des

autres. Dans la suite, IPF-Lasso1 fait référence à la méthode IPF-lasso avec la `pflist1` et IPF-Lasso2 à l'IPF-lasso avec la `pflist2`.

3.5.4 Critères d'évaluation

L'objectif principal de l'étude de simulation était d'évaluer et de comparer la capacité de sélection des différentes méthodes. Le premier critère d'évaluation était le taux de fausses découvertes des biomarqueurs (*False Discovery Rate* ou FDR) (GENOVESE et WASSERMAN [2002]), c'est-à-dire le taux de biomarqueurs inactifs sélectionnés (faux positifs FP) parmi tous les Q biomarqueurs sélectionnés (vrai positifs VP et FP).

$$FDR = \frac{FP}{FP + VP} = \frac{FP}{Q}.$$

Le second critère était le taux de faux négatifs (*False Negative Rate* ou FNR) (PAWITAN et al. [2005]), c'est-à-dire le taux de biomarqueurs qui ne sont pas sélectionnés par le modèle (faux négatifs FN) parmi ceux qui sont véritablement pronostiques (VP et FN).

$$FNR = \frac{FN}{FN + VP} = \frac{FN}{q}. \tag{3.25}$$

Ces critères varient entre 0 et 1 et sont basés sur le tableau de contingence 3.1 suivant :

TABLEAU 3.1 : La classification des biomarqueurs

	Actifs	**Inactifs**	Total
Sélectionnés	Vrai Positifs (VP)	Faux Positifs (FP)	Q
Non sélectionnés	Faux Négatifs (FN)	Vrai Négatifs (VN)	$p - Q$
Total	q	$p - q$	p

Dans notre étude de simulation, minimiser le FDR est le premier objectif, tandis qu'une faible valeur de FNR est une mesure complémentaire pour bien identifier les biomarqueurs actifs. À noter que pour le scénario 1, le nombre de biomarqueurs actifs est égal à 0 ($q = VP = FN = 0$). De fait, le FDR est égal à 0 si aucun biomarqueur n'est sélectionné ou 1 si au moins un biomarqueur est sélectionné.

Nous avons également mesuré le FDR et le FNR au niveau des groupes de biomarqueurs en étendant les définitions précédentes comme suit : un groupe est considéré comme sélectionné par une méthode si et seulement si la méthode sélectionnait au moins un biomarqueur appartenant à ce groupe. Nous avons défini le FDR des groupes comme étant le taux de groupes inactifs sélectionnés parmi tous les groupes sélectionnés et le FNR des groupes comme étant le taux de groupes actifs qui ne sont pas sélectionnés par la méthode parmi tous les groupes actifs.

En plus de ces deux mesures, nous avons calculé un autre critère exploratoire : la mesure $F1$ (*F-score*) qui résume à la fois le FDR et le FNR. Également connu sous le nom de coefficient de similarité de Dice (**dice1945measures**), il est défini comme la moyenne harmonique de la précision et de la sensibilité et varie entre 0 et 1. La mesure $F1$ est définie par :

$$F_1 = \frac{2VP}{2VP + FP + FN}. \tag{3.26}$$

Ainsi, les méthodes qui ont une bonne capacité de sélection selon ces critères sont celles qui se caractérisent par un faible FDR ainsi qu'un FNR aussi faible que possible. Ce qui se traduit par une valeur de $F1$ proche de 1.

3.5.5 Implémentation

L'implémentation des méthodes présentées précédemment a été réalisée avec le logiciel `R`. Les packages suivants ont été utilisés : `glmnet` (SIMON et al. [2011] et FRIEDMAN et al. [2018]) pour les méthodes lasso standard et lasso adaptatif ; `grpreg` (PATRICK [2020]) pour les méthodes cMCP et gel avec $\tau = 1/20$ comme suggéré par BREHENY [2015] pour effectuer une sélection des groupes et à l'intérieur des groupes. `ipflasso` (BOULESTEIX et FUCHS [2015]) pour le IPF-Lasso ; `SGL` (SIMON et al. [2019]) pour la méthode SGL avec la valeur par défaut du paramètre $\alpha = 0,95$; `corpcor` (SCHAFER et al. [2017]) pour faire l'analyse en composantes principales. Enfin, `timeROC` (BLANCHE et al. [2013], BLANCHE [2019]) a été utilisé pour estimer l'aire sous la courbe ROC (AUC) en fonction du temps.

3.5.6 Résultats

Résultats de l'étude de simulation (a) avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale

Évaluation de la sélection au niveau des biomarqueurs

Le tableau B.2 de l'annexe B et les figures 3.2 et 3.3 résument les résultats de FDR, FNR et le nombre de biomarqueurs sélectionnés par les différentes méthodes sur les 500 répétitions.

Scénario nul 1 ($l = 0$ et $q = 0$).

Le FNR n'est pas calculable et tous les biomarqueurs sélectionnés étaient des faux positifs, ce qui donne un FDR moyen égal à la probabilité qu'un modèle sélectionne au moins un biomarqueur. Les méthodes cMCP, gel, SGL, IPF-Lasso1 et IPF-Lasso2 avaient un FDR et un nombre moyen de biomarqueurs sélec-

tionnés inférieur aux méthodes lasso standard et lasso adaptatif. En effet, le FDR variait de 0, 25 à 0, 64 et le nombre de biomarqueurs sélectionnés était entre 0, 98 et 8, 99 pour ces méthodes. D'autre part, le lasso standard et le lasso adaptatifs avaient un FDR égal à 1 et des valeurs plus élevées de nombre de biomarqueurs sélectionnés, à l'exception de la pondération $Lasso + PCA$ (8, 27). Cela signifie que ces méthodes sélectionnent au moins un biomarqueur pour chaque jeu de données.

Scénario nul 2 ($l = 20$ et $q = 100$).
Rappelons que pour ce scénario les 100 biomarqueurs actifs sont répartis équitablement sur les 20 groupes actifs supprimant tout effet de groupe. Les méthodes gel, cMCP, lasso adaptatif avec les pondérations SW, $ASW \times SW$ et $MSW \times SW$ avaient un FDR plus faible que les autres méthodes avec des valeurs moyennes, respectivement, de 0,05, 0,10, 0,13, 0,17, 0,18. Toutefois, le nombre moyen de biomarqueurs sélectionnés par les méthodes de la cMCP et gel était inférieur à celui des autres méthodes, soit, respectivement, 35,01 et 50,68. Alors que, les pondérations SW, $ASW \times SW$ et $MSW \times SW$ présentaient un nombre moyen de biomarqueurs sélectionnés assez similaire : 97, 91 , 98, 36 et 95, 77, respectivement.

Scénarios alternatifs 3 ($l = 1$ et $q = 8$) et 4 ($l = 2$ et $q = 8$).
Toutes les méthodes ont réduit le FDR par rapport au lasso standard, qui présentait un FDR de 0, 81. Les résultats montrent que les méthodes optimales étaient le lasso adaptatif avec les pondérations $ASW \times SW$ et $MSW \times SW$. Elles avaient les plus faibles valeurs de FDR/FNR et un nombre moyen de biomarqueurs sélectionnés proche du nombre de biomarqueurs réellement actifs ($q = 8$), soit 0, 06/0, 02 (8, 35) pour le scénario 3 ; 0, 11/0, 03 (8, 87) et 0, 11/0, 03 (8, 80), respectivement, pour le scénario 4. D'autre part, la méthode cMCP augmentait considérablement le FNR par rapport au lasso standard de 0, 04 à 0, 58 dans le

scénario 3 et de 0,04 à 0,60 dans le scénario 4. La méthode gel avait des valeurs de FDR/FNR relativement faibles mais supérieures aux valeurs obtenues par les pondérations $ASW \times SW$ et $MSW \times SW$. Dans le scénario 4, avec $l = 2$ groupes actifs, toutes les méthodes sélectionnaient plus de biomarqueurs et avaient un FNR plus élevé par rapport au scénario 3 ($l = 1$).

Scénarios alternatifs 5 ($l = 1$ et $q = 32$) et 6 ($l = 2$ et $q = 32$).
Les résultats montrent que toutes les méthodes avaient un FDR plus faible par rapport au lasso standard. Dans le scénario 5, les méthodes gel et lasso adaptatif avec les pondérations ASW, MSW, $ASW \times SW$ et $MSW \times SW$ présentaient le meilleur équilibre FDR-FNR. Pour ces pondérations, le nombre de biomarqueurs sélectionnés était légèrement inférieur au nombre de biomarqueurs réellement actifs ($q = 32$), tandis que, la méthode gel avait un nombre de biomarqueurs sélectionnés (43, 72) supérieur à celui-ci ($q = 32$). Dans le scénario 6, les pondérations $ASW \times SW$ et $MSW \times SW$ présentaient le meilleur équilibre FDR-FNR et étaient également les plus parcimonieuses en termes de nombre de biomarqueurs sélectionnés. À l'inverse, la méthode gel avait le FDR le plus faible mais elle augmentait le FNR avec une valeur de 0, 80 par rapport au lasso standard avec 0, 50. De plus, cette méthode avait un nombre de biomarqueurs sélectionnés inférieur au nombre de biomarqueurs réellement actifs. D'autre part, dans les scénarios 5 et 6, les méthodes SGL et le lasso adaptatif avec les pondérations PCA et $Lasso + PCA$ ont montré les plus mauvais résultats.

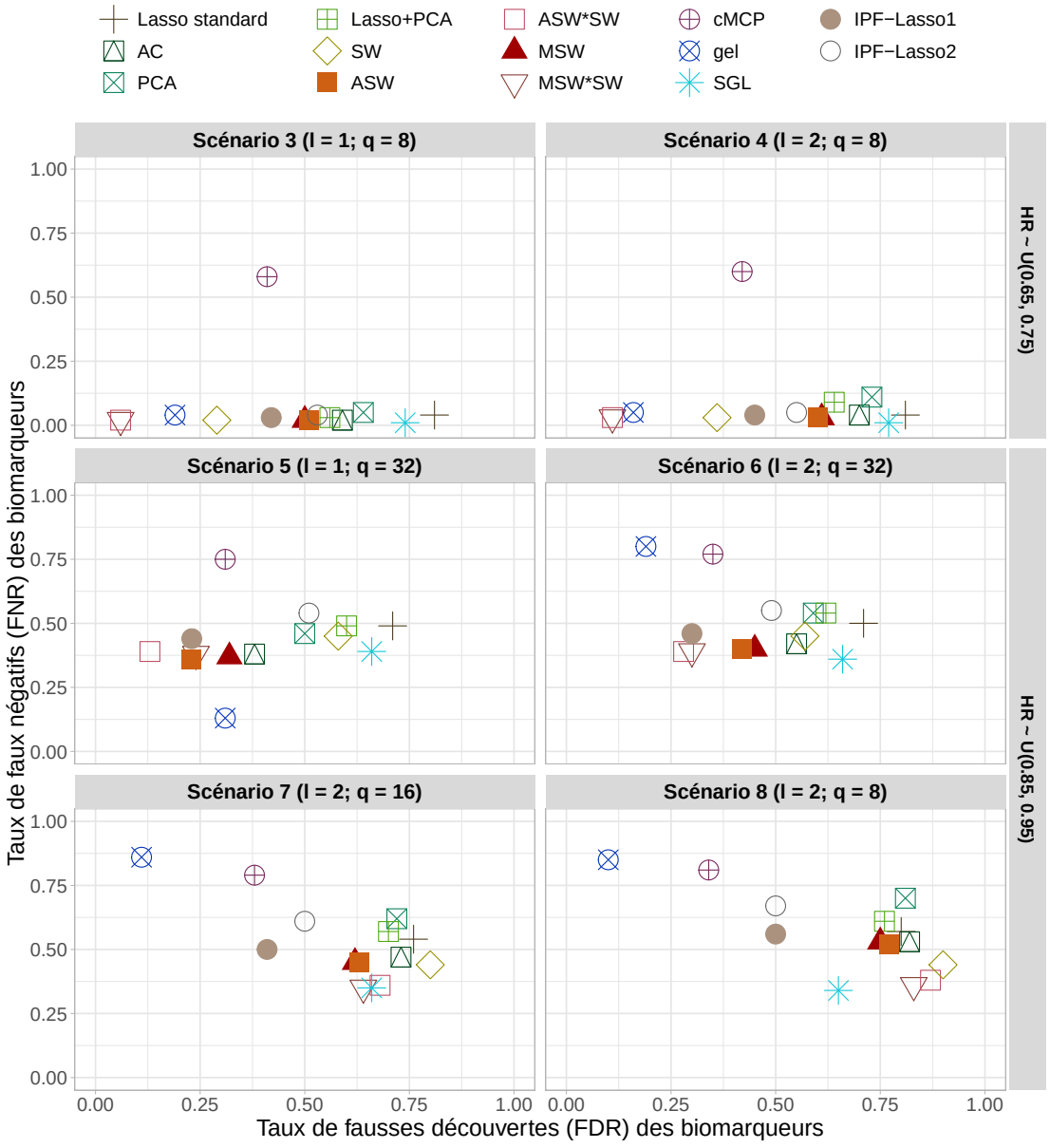
Scénarios alternatifs 7 ($l = 2$ et $q = 16$) et 8 ($l = 2$ et $q = 8$).
Pour ces scénarios, les résultats étaient assez similaires pour la plupart des méthodes. Les méthodes cMCP et gel avaient le FDR le plus faible avec 0, 38 et 0, 11, respectivement, $versus$ 0, 76 pour le lasso standard dans le scénario 7 et 0, 34 et 0, 10, respectivement, $versus$ 0, 80 pour le lasso standard dans le scénario 8. Néanmoins, ces deux méthodes ont fortement augmenté le FNR avec

0,79 et 0, 86, respectivement, $versus$ 0, 54 pour le lasso standard dans le scénario 7 et 0, 81 et 0, 85, respectivement, $versus$ 0, 57 pour le lasso standard dans le scénario 8. De plus, elles ont également sélectionné en moyenne moins de biomarqueurs que le nombre de biomarqueurs réellement actifs dans les deux scénarios ($q = 16$ et $q = 8$). La méthode IPF-Lasso1 a fourni le meilleur compromis entre FDR et FNR : 0, 41 et 0, 50 dans le scénario 7, et 0, 50 et 0, 56 dans le scénario 8. Le lasso adaptatif avec la pondération PCA a montré les plus mauvais résultats avec des valeurs FDR/FNR de 0, 81 et 0, 70.

Dans les scénarios alternatifs 3-6, la méthode IPF-Lasso avec la liste de facteurs de pénalité 2 (`pflist2`) avait de meilleures performances que celles obtenues avec les facteurs de pénalisation 1 (`pflist1`) en termes d'équilibre FDR-FNR. À noter que, dans les scénarios 1, 7 et 8, la méthode gel avait 268, 1 et 26 itérations qui n'ont pas convergé, respectivement. Toutes les autres méthodes n'avaient pas de problèmes de convergence.

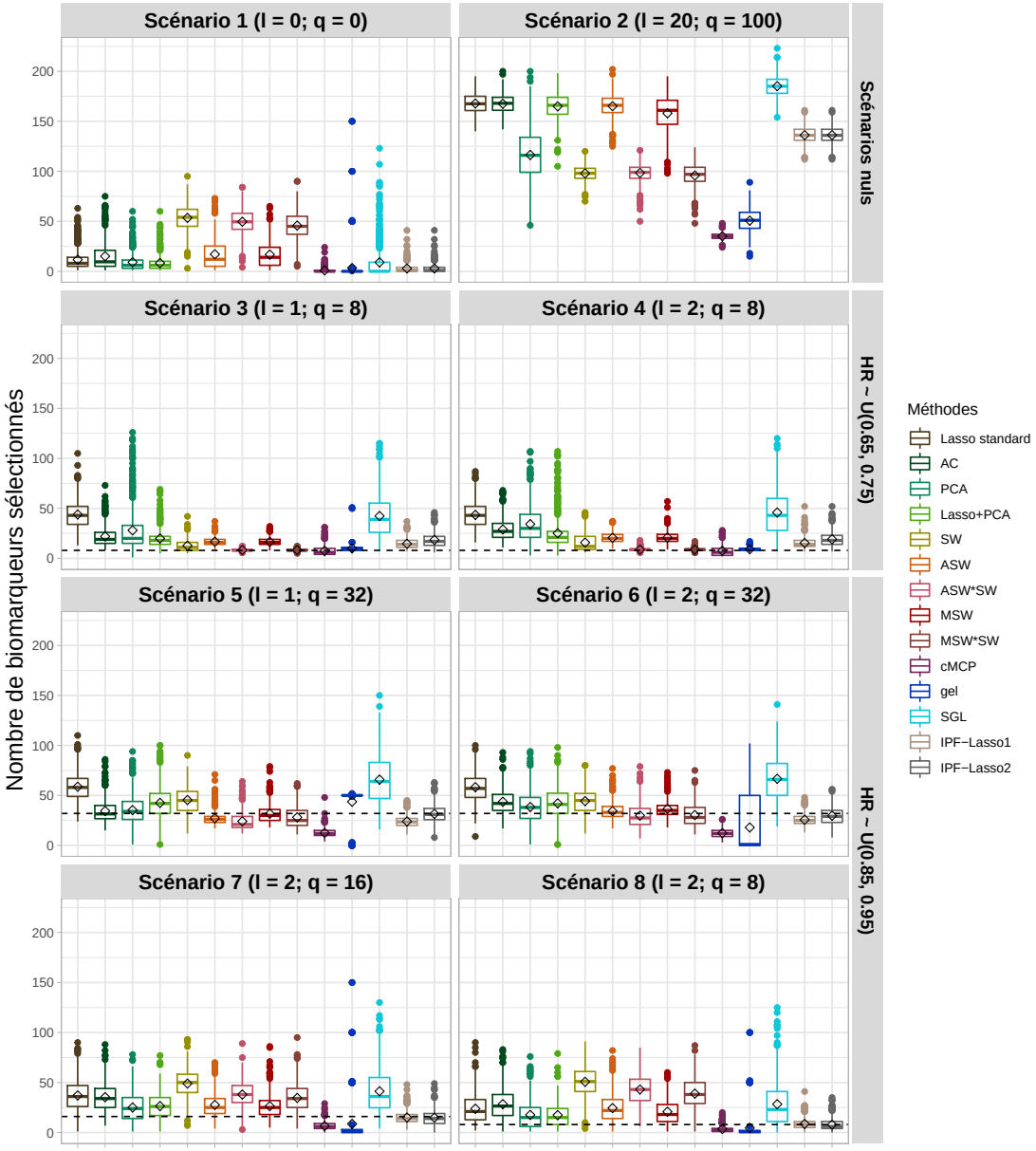
Globalement, dans tous les scénarios, les approches de pondération MSW et $MSW \times SW$ semblent donner les meilleurs résultats, avec le moins de biomarqueurs sélectionnés, la plus forte réduction du FDR, et une réduction ou une augmentation non significative du FNR.

FIGURE 3.2 : *Résultats de la simulation (a) au niveau des biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE 3.3 : *Résultats de la simulation (a) au niveau des biomarqueurs. Boxplots du nombre de biomarqueurs sélectionnés dans tous les scénarios*



*Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. La boîte délimite l'intervalle interquartile (IQR) et contient une ligne horizontale correspondant à la médiane ; en dehors de la boîte, les moustaches de style Tukey s'étendent jusqu'à un maximum de 1,5*IQR au-delà de la boîte. Les losange noirs représentent le nombre moyen de biomarqueurs sélectionnés. Les lignes noires en pointillés indiquent le nombre de biomarqueurs réellement actifs dans les scénarios alternatifs. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée*

Évaluation de la sélection au niveau des groupes de biomarqueurs

Le tableau B.3 et la figure B.1 de l'annexe B et la figure 3.4 résument les résultats de FDR, FNR et du nombre de groupes de biomarqueurs sélectionnés par chaque méthode sur 500 répétitions.

Scénario nul 1 ($l = 0$ et $q = 0$).

Comme dit précédemment, le FNR n'est pas calculable dans ce scénario et tous les groupes sélectionnés étaient des faux positifs, ce qui donne un FDR moyen égal à la probabilité qu'une méthode sélectionne au moins un groupe. Les méthodes cMCP, gel, SGL, IPF-Lasso1 et IPF-Lasso2 avaient un FDR et un nombre de groupes sélectionnés inférieur aux méthodes lasso standard et lasso adaptatif. Par exemple, le FDR et le nombre de groupes sélectionnés étaient de 0, 25 et 0, 32 pour la méthode gel et de 0,64 et 1, 28 pour les deux méthodes IPF-Lasso1 et IPF-Lasso2 $versus$ 1 et 7, 55 pour la méthode lasso standard.

Scénario nul 2 ($l = 20$ et $q = 100$).

Tous les groupes sont actifs ainsi le FDR n'est pas calculable dans ce scénario. Dans l'ensemble, toutes les méthodes ont sélectionné tous les groupes actifs, à l'exception de la pondération PCA, qui avait le FNR le plus élevé 0, 21 et un nombre de groupes sélectionnés de 15, 78. Aussi, la méthode gel avait un FNR de 0,04 et un nombre de groupes sélectionnés de 19,11.

Scénarios alternatifs 3 ($l = 1$ et $q = 8$) et 4 ($l = 2$ et $q = 8$).

Les méthodes gel et lasso adaptatif avec les pondérations ASW, $ASW \times SW$, MSW, $MSW \times SW$ avaient un FDR et un FNR presque égal à 0. Elles étaient également les plus parcimonieuses en termes de nombre de groupes sélectionnés, ce qui signifie que ces méthodes ont souvent sélectionné uniquement des biomarqueurs appartenant aux groupes actifs.

Scénarios alternatifs 5 ($l = 1$ et $q = 32$) et 6 ($l = 2$ et $q = 32$).

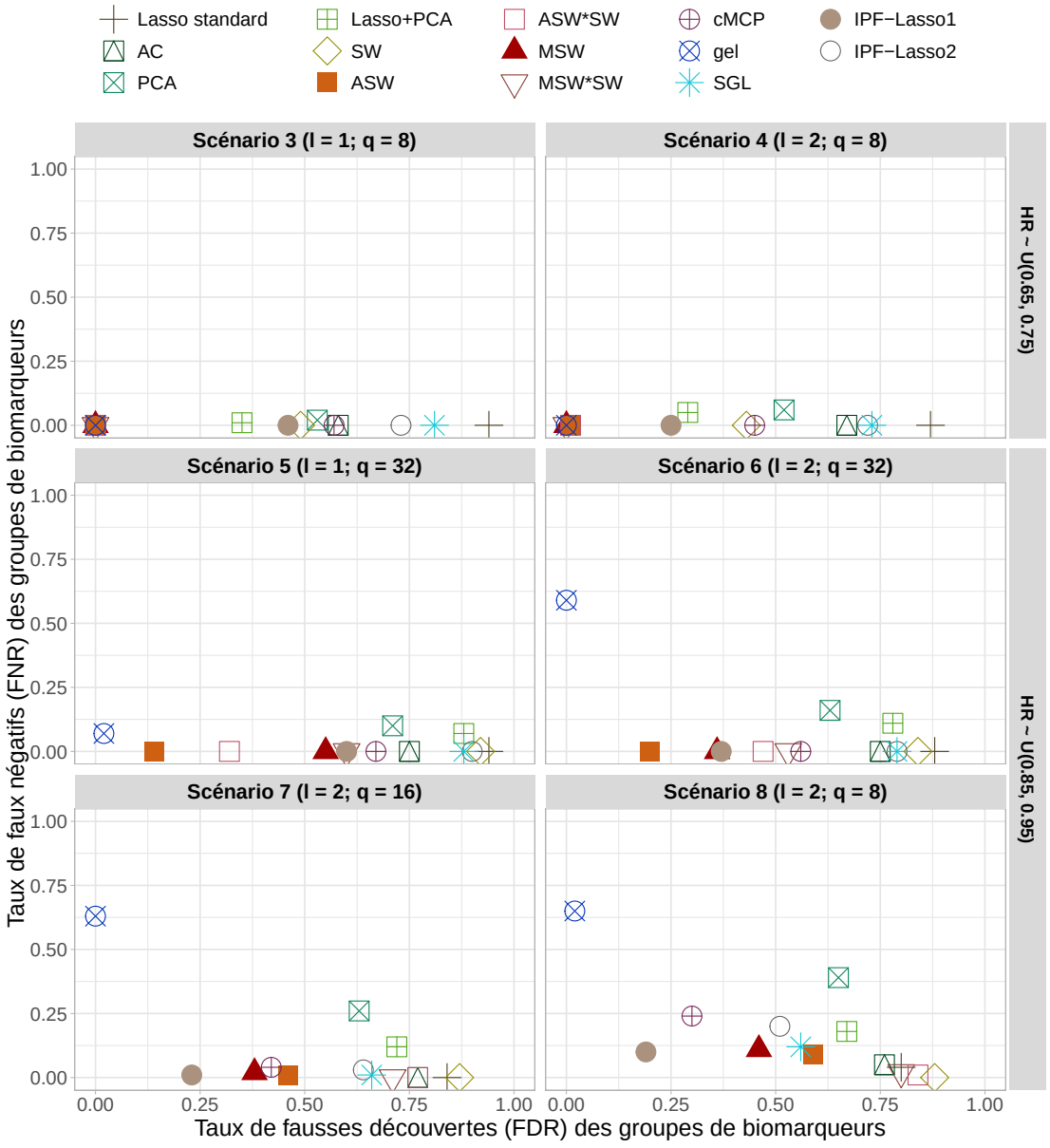
Les résultats montrent que toutes les méthodes augmentaient le FDR par rap-

port aux scénarios 3 et 4. Néanmoins, les pondérations ASW, $ASW \times SW$, MSW, $MSW \times SW$ ont continué à obtenir de meilleurs résultats en termes de FDR, FNR et de nombre de groupes sélectionnés par rapport aux autres. La méthode gel avait le FDR le plus faible mais le FNR le plus élevé $(0,59)$ dans le scénario 6.

Scénarios alternatifs 7 ($l = 2$ et $q = 16$) et 8 ($l = 2$ et $q = 8$).
Le FDR, le FNR et le nombre de groupes sélectionnés ont augmenté pour toutes les méthodes. La méthode IPF-lasso1 présentait le meilleur équilibre FDR-FNR. La méthode gel avait le FNR le plus élevé, soit 0,65, et le nombre de groupes sélectionnés inférieur au nombre de groupes réellement actifs ($l = 2$).

Nous avons, également, représenté les résultats des FDR, FNR et de la mesure $F1$ avec des *heatmaps* accompagnés de statistiques descriptives (médiane, minimum, et maximum) (voir annexe B, figures B.6-B.11). Les résultats montrent globalement que le lasso adaptatif avec double poids au niveau individuel et au niveau du groupe, c'est-à-dire les pondérations $MSW \times SW$ et $ASW \times SW$, était parmi les méthodes les plus performantes en termes de balance FDR-FNR et du score $F1$. Néanmoins, la méthode IPF-lasso avec la liste de facteurs de pénalité 1 (`pflist1`) avait les valeurs les plus élevées du score $F1$ dans les scénarios 7 et 8 par rapport aux autres méthodes. Les méthodes gel et cMCP avaient, en général, un FDR très faible mais un FNR sensiblement élevé, ceci résultait en un score $F1$ se situant dans la moyenne en comparaison avec les autres méthodes. À l'inverse, la méthode SGL avait un faible FNR en général, mais un FDR relativement élevé et un score $F1$ faible.

FIGURE 3.4 : *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des groupes de biomarqueurs dans les scénarios alternatifs*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

Évaluation de la capacité prédictive

Le tableau 3.2 résume les valeurs moyennes de l'AUC à 5 ans et l'erreur type empirique (ETE) sur les 500 répétitions. Globalement, toutes les méthodes avaient des résultats similaires dans tous les scénarios alternatifs. En particulier, les méthodes $MSW \times SW$ et $ASW \times SW$ avaient l'AUC le plus élevé, tandis que, la méthode gel avait l'AUC le plus faible.

Résultats de l'étude de simulation (b) avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de tailles différentes (10 de taille 25 et 10 de taille 75)

Les résultats de l'étude de simulation (b) étaient cohérents avec les résultats de l'étude de simulation (a). Généralement, les pondérations ASW, $ASW \times SW$, MSW, $MSW \times SW$ avaient le meilleur équilibre FDR-FNR, sauf pour les scénarios alternatifs 7 et 8, dans lesquels l'IPF-Lasso1 a montré de meilleurs résultats avec des valeurs plus faible de FDR et de FNR en comparant aux autres méthodes (voir tableau B.4 et figures B.2 et B.3 de l'annexe B). Nous avons observé que la méthode gel avait un faible FDR des biomarqueurs (et groupes) mais qu'elle augmentait le FNR des biomarqueurs (et groupes) par rapport au lasso standard et sélectionnait généralement peu de biomarqueurs (et groupes) (voir tableau B.5 et figures B.4 et B.5 de l'annexe B).
Le tableau B.6 de l'annexe B montre également que toutes les méthodes ont fourni des valeurs moyennes de l'AUC à 5 ans assez similaires.

Pareillement, les *heatmaps* des FDR, FNR et $F1$ (figures B.12-B.17 de l'annexe B) ont montré des résultats semblables à ceux de l'étude de simulation (a). La méthode lasso adaptatif avec les pondérations $MSW \times SW$ et $ASW \times SW$

avait la meilleur balance FDR-FNR et des valeurs élevées du score $F1$ dans la plupart des scénarios sauf les scénarios 7 et 8, où la méthode IPF-lasso1 avait des valeurs de la mesure $F1$ les plus élevées.

Comme pour l'étude de simulation (a) et les scénarios 1, 7 et 8, la méthode gel n'a pas convergé pour 327, 6 et 27 jeux de données, respectivement.

Résultats de l'étude de sensibilité

Avec une étude de sensibilité dans un contexte à très haute dimension, nous avons étudié un cas plus extrême en utilisant le même ensemble de données de simulation (a) avec le même nombre de biomarqueurs $p = 1000$ mais seulement avec une taille d'échantillon $n = 50$. Dans cette étude, nous avons évalué la capacité de sélection des méthodes lasso standard et lasso adaptatif avec les différentes pondérations. Les résultats présentés dans le tableau B.9 de l'annexe B montre que, globalement, les valeurs de FDR et FNR ont augmenté par rapport à l'étude de simulation (a), comme attendu. Néanmoins, dans les scénarios alternatifs, la méthode lasso adaptatif avec la pondération $MSW \times SW$ a réduit le FDR par rapport au lasso standard.

TABLEAU 3.2 : *Résultats de la simulation (a). La moyenne et l'erreur standard empirique (ESE) de l'AUC à 5 ans*

Méthodes	**Scénario 3** ($l = 1$; $q = 8$)	**Scénario 4** ($l = 2$; $q = 8$)	**Scénario 5** ($l = 1$; $q = 32$)	**Scénario 6** ($l = 2$; $q = 32$)	**Scénario 7** ($l = 2$; $q = 16$)	**Scénario 8** ($l = 2$; $q = 8$)
	HR$\sim U(0,65,\ 0,75)$		HR$\sim U(0,85,\ 0,95)$			
Lasso standard	0.88 (0.02)	0.87 (0.03)	0.75 (0.04)	0.75 (0.04)	0.67 (0.05)	0.61 (0.05)
Lasso adaptatif						
AC	0.88 (0.02)	0.88 (0.03)	0.78 (0.04)	0.77 (0.04)	0.68 (0.05)	0.61 (0.05)
PCA	0.87 (0.06)	0.85 (0.05)	0.74 (0.09)	0.73 (0.07)	0.64 (0.07)	0.58 (0.06)
$Lasso + PCA$	0.88 (0.04)	0.86 (0.05)	0.74 (0.08)	0.73 (0.06)	0.66 (0.06)	0.61 (0.06)
SW	0.88 (0.02)	0.88 (0.03)	0.75 (0.04)	0.75 (0.04)	0.65 (0.05)	0.59 (0.05)
ASW	0.88 (0.02)	0.88 (0.03)	0.78 (0.03)	0.77 (0.04)	0.68 (0.05)	0.61 (0.05)
$ASW \times SW$	0.89 (0.02)	0.88 (0.02)	0.78 (0.04)	0.77 (0.04)	0.66 (0.05)	0.59 (0.05)
MSW	0.88 (0.02)	0.88 (0.02)	0.78 (0.04)	0.77 (0.04)	0.68 (0.05)	0.62 (0.05)
$MSW \times SW$	0.89 (0.02)	0.88 (0.02)	0.77 (0.04)	0.76 (0.04)	0.67 (0.05)	0.60 (0.05)
cMCP	0.87 (0.03)	0.86 (0.03)	0.75 (0.04)	0.74 (0.04)	0.67 (0.05)	0.61 (0.06)
gel	0.88 (0.02)	0.88 (0.02)	0.74 (0.08)	0.60 (0.08)	0.58 (0.07)	0.57 (0.06)
SGL	0.88 (0.02)	0.88 (0.02)	0.77 (0.04)	0.76 (0.04)	0.69 (0.05)	0.62 (0.06)
IPF-Lasso						
IPF-Lasso1	0.88 (0.02)	0.88 (0.03)	0.78 (0.04)	0.77 (0.04)	0.69 (0.05)	0.63 (0.05)
IPF-Lasso2	0.88 (0.02)	0.88 (0.03)	0.76 (0.04)	0.75 (0.04)	0.68 (0.05)	0.62 (0.06)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9))).

IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9))).

3.6 Application : cancer du sein précoce

Contexte

Les méthodes présentées ont également été appliquées au jeu de données réelles, présenté précédemment (3.2.2). Il est issu de plusieurs essais cliniques et constitué de 614 patientes atteintes d'un cancer du sein précoce et pour lesquelles des données d'expression de gènes sont disponibles. Les patientes ont été traitées par chimiothérapie adjuvante (postopératoire) à base d'anthracycline avec ou sans l'ajout de taxanes. Le taux de survie à 5 ans sans rechute chez ces patientes est estimé à 74% [95% IC : 69% -77%] et le taux de censure est estimé à de 78% (134 événements). L'ensemble des données sur l'expression génique des patientes est disponible dans package `biospear`.
Quatre voies ont été choisies à partir d'un travail antérieur sur le rôle pronostique des voies biologiques dans le cancer du sein précoce (IGNATIADIS et al. [2012]). Les voies ont été modélisées comme des moyennes pondérées des gènes inclus : 3 voies avec un effet pronostique (système immunitaire, prolifération et invasion stroma) et 1 sans effet pronostique (voie d'activation SRC) servant de voie de contrôle. Un ensemble de 128 gènes standardisés de ce jeu de données d'application est réparti sur ces voies disjointes comme suit : "**Système immunitaire**" (53 gènes), "**Prolifération**" (44 gènes), "**Stroma**" (19 gènes) et "**SRC**" (12 gènes).

L'objectif de cette application est d'identifier des gènes potentiellement pronostiques appartenant à des voies présumées jouer un rôle pronostique dans le cancer du sein.
Pour la méthode IPF-Lasso, une liste de facteurs de pénalité a été fixée `pflist` pénalisant un ou deux groupes deux fois plus ou moins que les autres :

```
pflist = list(
```

```
  c(1, 2, 2, 2),
  c(2, 1, 2, 2),
  c(2, 2, 1, 2),
  c(2, 2, 2, 1),
  c(2, 1, 1, 1),
  c(1, 2, 1, 1),
  c(1, 1, 2, 1),
  c(1, 1, 1, 2),
  c(1, 1, 2, 2),
  c(1, 2, 1, 2),
  c(1, 2, 2, 1),
  c(2, 1, 1, 2),
  c(2, 1, 2, 1),
  c(2, 1, 2, 1),
  c(2, 2, 1, 1)
)
```

Les données d'expression de gènes ont été divisées 500 fois aléatoirement en échantillon d'apprentissage (60%) et échantillon de validation (40%). Les méthodes de sélection des gènes ont été appliquées à chaque ensemble d'apprentissage et l'AUC a été calculé pour chaque ensemble de validation associé.

Résultats

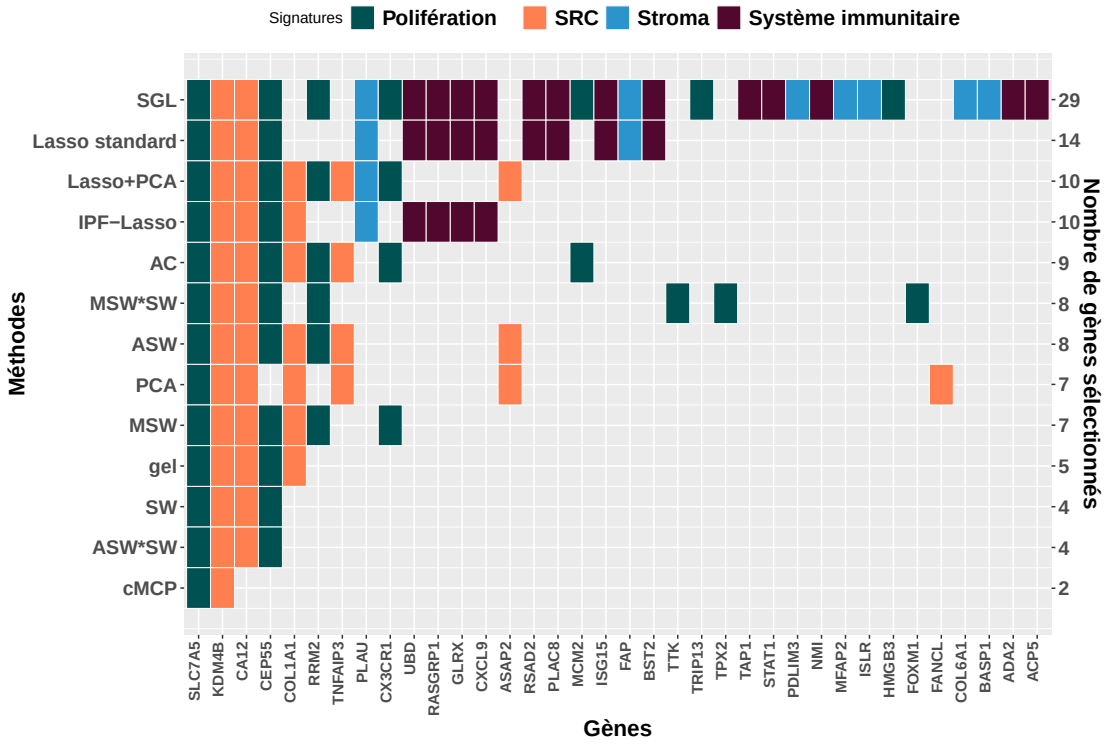
La figure 3.5 présente les gènes identifiés dans au moins 10% des 500 répétitions, ainsi que le nombre de gènes sélectionnés par chaque méthode. Les résultats montrent que les gènes appartenant au groupe Prolifération ont été sélectionnés plus fréquemment par plusieurs méthodes que les groupes SRC, Système immunitaire et Stroma. Nous avons observé une certaine variation

entre les méthodes en ce qui concerne le nombre de biomarqueurs sélectionnés : la méthode SGL a fourni le modèle le moins parcimonieux, avec 29 gènes sélectionnés (2 dans le groupe SRC, 7 dans le groupe Stroma, 7 dans le groupe Prolifération, 13 dans le groupe Système immunitaire). La méthode lasso standard a sélectionné 14 gènes (2 en SRC, 2 en Stroma, 2 en Prolifération, 8 en Système Immunitaire). La méthode IPF-Lasso a sélectionné 10 gènes (1 en Stroma, 3 en SRC, 2 en Prolifération, 4 en Système Immunitaire). Le cMCP n'a sélectionné que 2 gènes appartenant aux signatures SRC et Prolifération (SLC7A5, KDM4B). La méthode gel a sélectionné 5 gènes (2 dans la Prolifération ; 3 dans le SRC) (SLC7A5, KDM4B, CA12, CEP55, COL1A1).
La méthode proposée, lasso adaptatif avec pondération MSW, a sélectionné 7 gènes (3 en SRC ; 4 en Prolifération) alors que la pondération ASW a sélectionné 8 gènes (3 en Prolifération ; 5 en SRC). La pondération $MSW \times SW$ a sélectionné 8 gènes dont 2 gènes appartenant au SRC (KDM4B, CA12) et 6 appartenant à la signature Prolifération, précédemment suggérée comme étant la voie la plus pronostique du cancer du sein précoce (Haibe-Kains et al. [2008]) (SLC7A5, CEP55, RRM2, TTK, TPX2, FOXM1). Enfin, les pondérations SW et $ASW \times SW$ ont sélectionné 4 gènes dont 2 gènes appartenant au SRC (KDM4B, CA12) et 2 appartenant à la signature sur la prolifération (SLC7A5, CEP55).

Le tableau 3.3 présente la moyenne et l'erreur type empirique (ETE) de l'AUC, évaluant la capacité de prédiction de la survie sans rechute à 5 ans sur les 500 ensembles de validation.
La méthode de référence, lasso standard, avait une AUC moyenne de 0,61 avec une ETE de 0,04. Cependant, une faible différence a été observée entre les valeurs moyennes des AUC des différentes méthodes. La méthode lasso adaptatif avec les pondérations $Lasso + PCA$ et PCA a donné l'AUC le plus élevée 0,63 (ETE 0,04). La méthode SGL a fourni une AUC de 0,62 (ETE 0,04), et le lasso

adaptatif avec pondération MSW avait une AUC de 0,61 (ETE 0,04). La méthode gel a donné l'AUC la plus faible, soit 0,57 (ETE 0,05), et la méthode du cMCP avait une AUC de 0,60 (ETE 0,05). Le lasso adaptatif avec les pondération AC, ASW, SW, $MSW \times SW$ et $ASW \times SW$ avait une valeur moyenne d'AUC similaire au lasso standard, soit 0,62 (ETE 0,04).

FIGURE 3.5 : *Les biomarqueurs identifiés par les méthodes sur 500 échantillons d'apprentissage des donnés de l'expression de gènes du cancer du sein pour la survie sans rechute*



L'axe des x représente les gènes identifiés dans au moins 10% des 500 répétitions par chaque méthode. AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

TABLEAU 3.3 : *La moyenne et l'erreur type empirique (ETE) de l'AUC à 5 ans pour la prédiction de la survie sans rechute obtenues à partir de 500 échantillons de validation des donnés de l'expression de gènes du cancer du sein*

Méthodes	AUC	ETE
SGL	0.62	0.04
Lasso standard	0.61	0.04
$Lasso + PCA$	0.63	0.04
IPF-Lasso	0.61	0.05
AC	0.62	0.04
$MSW \times SW$	0.62	0.04
ASW	0.62	0.04
PCA	0.63	0.04
MSW	0.61	0.04
gel	0.57	0.05
SW	0.62	0.04
$ASW \times SW$	0.62	0.04
cMCP	0.60	0.05

AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

3.7 Discussion

Dans le contexte de la médecine de précision et des données de haute dimension, un problème commun est de savoir comment prendre en compte des voies biologiques (*pathways*) ou des groupes de biomarqueurs dans la procédure de sélection des variables. La régression lasso est une méthode standard de sélection de biomarqueurs à partir d'un ensemble de candidats de haute dimension. Néanmoins, il existe à ce jour peu de méthodes qui tiennent compte de la structure en groupe dans la sélection lasso. Dans ce travail de thèse, nous avons proposé différentes stratégies de pondération destinées à la méthode lasso adaptatif afin d'identifier les biomarqueurs actifs appartenant à des voies ou groupes disjoints préspécifiés dans le cadre de la survie. Ces pondérations favorisent la sélection de biomarqueurs appartenant aux groupes les plus pronostiques. La sélection des biomarqueurs peut être améliorée en exploitant les informations au niveau du groupe.

Dans les études de simulation, nous avons comparé l'approche que nous proposons, basée sur le lasso adaptatif, aux méthodes cMCP, gel, SGL et IPF-Lasso en termes de sélection de variables et de performances de prédiction. Sur la base des résultats de l'étude de simulation, les méthodes de pondération du lasso adaptatif avec le produit $MSW \times SW$ et le produit $ASW \times SW$ surpassent leurs concurrents en réduisant le FDR sans augmenter le FNR de façon substantielle (équilibre FDR-FNR), tant en termes de biomarqueurs qu'en termes de groupes. En outre, ces méthodes sont les plus parcimonieuses en termes de nombre de biomarqueurs sélectionnés. De plus, en termes de performance de prédiction, les pondération $MSW \times SW$ et $ASW \times SW$ avaient souvent la valeur moyenne d'AUC la plus élevée.

Bien que les méthodes gel et cMCP aient montré un faible FDR, ils avaient un

FNR plus élevé que le lasso standard. Globalement, la méthode SGL a donné de très mauvais résultats et a sélectionné à tort un grand nombre de biomarqueurs. À l'inverse, les méthodes MCP et gel ont souvent fourni des solutions parcimonieuses. Cependant, en termes de performance de prédiction, la méthode gel a montré de moins bonne performance que les autres méthodes. Même si la méthode IPF-Lasso a montré de bonnes performances dans certains scénarios alternatifs, cette approche nécessite de préspécifier arbitrairement la liste des facteurs de pénalités candidats à tester, ce qui exige une connaissance, a priori, assez solide sur les rôles pronostiques possibles des différents groupes. À mesure que le nombre de groupes augmente, cette limitation peut devenir critique et l'impact du choix des facteurs de pénalités sur le résultat peut être important et ceci avec un temps de calcul extrêmement important.

Nous avons utilisé, dans les études de simulation, les valeurs par défaut des paramètres α et τ des méthodes SGL et gel, respectivement. Afin de vérifier que cela ne désavantage pas ces méthodes, nous avons effectué une analyse de sensibilité sur le même ensemble de données de simulation (a) en utilisant une boucle de validation croisée supplémentaire pour le choix de ces paramètres de réglage sur une grille de valeurs. Les résultats de ces analyses présentés dans les tableaux B.7 et B.8 de l'annexe B montrent qu'il n'y a pas de différence importante entre les résultats obtenus par SGL avec l'utilisation de la validation croisée pour le choix de α, et les résultats obtenus par SGL avec la valeur par défaut de α, présentés dans les tableaux 2 et 3 de l'annexe B. Ces tableaux montrent aussi que la méthode gel, avec une valeur de τ choisie par validation croisée, a donné des résultats moins bons (avec un FDR plus élevé, un nombre de biomarqueurs sélectionnés et un nombre de groupes sélectionnés) que ceux avec la valeur fixée par défaut de τ.

Récemment, Tang et al. [2019] ont proposé une méthode bayésienne appelée *Gsslasso Cox*. Il s'agit d'un modèle hiérarchique bayésien permettant de prédire les résultats de survie de la maladie et détecter les gènes associés en incorporant des structures en groupe de voies biologiques. Ils ont comparé leur méthode avec des alternatives fréquentistes telles que le lasso standard, le SGL et le cMCP. Dans leur étude de simulation, ils ont constaté que le cMCP était systématiquement meilleur que les autres approches fréquentistes. En revanche, dans notre étude de simulation, le cMCP s'est avéré globalement moins performant que le lasso adaptatif avec la pondération $MSW \times SW$ que nous proposons. De même, bien que nous n'avons envisagé que des approches fréquentistes, les connaissances a priori sur les groupes de biomarqueurs pourraient correspondre à une approche bayésienne (L. Zhang et al. [2014]) avec une méthode de sélection hiérarchique structurée de variables (*hierarchical structured variable selection* ou HSVS).

Dans ce deuxième axe de thèse, nous nous sommes limités à des méthodes de sélection de variables basées sur le lasso avec le modèle de Cox. D'autres approches comme le *component-wise boosting* (Binder et Schumacher [2009]), ou le *Group and sparse group partial least square (gPLS, sgPLS)* (Liquet et al. [2016]) pourraient être envisagées, mais ne relèvent pas du champ d'application de ce travail de thèse.

Enfin, nous nous sommes concentrés sur les groupes de biomarqueurs non corrélés et disjoints, mais les méthodes présentées peuvent être étendues au cas des groupes qui se chevauchent (c'est-à-dire avec des biomarqueurs appartenant à plusieurs groupes à la fois) par duplication de biomarqueurs (Obozinski et al. [2011]). D'autre part, nous avons considéré dans ce travail l'exemple des gènes appartenant à des voies fonctionnelles, néanmoins, les mêmes méthodes peuvent être utilisées dans un éventail plus large d'applications, comme le cas

des ensembles de données génomiques de nature et de sources différentes.

3.8 Conclusion

Lors de l'ajustement de modèles de régression de Cox pénalisés en haute dimension, la sélection des biomarqueurs peut prendre en compte leur appartenance à des groupes préspécifiés ayant un rôle biologique supposé. Nous avons proposé une approche basée sur le lasso adaptatif, avec différentes stratégies de pondération, y compris celles basées sur les statistiques du test de Wald. Nous préconisons en particulier l'utilisation du schéma de pondération $MSW \times SW$ qui, en pondérant différemment chaque biomarqueur, prend en compte à la fois le rôle pronostique des groupes de biomarqueurs et le rôle pronostique individuel de chaque biomarqueur. Dans nos simulations, cette méthode a présente la meilleur balance FDR-FNR par rapport aux autres méthodes.

Ce deuxième axe de travail de thèse a abouti à un article publié dans la revue scientifique *BMC Bioinformatics* : "Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models".

3.9 Références

SLAMON, D. J., CLARK, G. M., WONG, S. G., LEVIN, W. J., ULLRICH, A. & MCGUIRE, W. L. (1987). Human breast cancer : correlation of relapse and survival with amplification of the HER-2/neu oncogene. *science*, *235*(4785), 177-182. https://doi.org/10.1126/science.3798106. 59

ROA, I., de TORO, G., SCHALPER, K., de ARETXABALA, X., CHURI, C. & JAVLE, M. (2014). Overexpression of the HER2/neu gene : a new therapeutic possibility for patients with advanced gallbladder cancer. *Gastrointestinal cancer research : GCR*, *7*(2), 42. 59.

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. 61

VERWEIJ, P. J. M. & HOUWELINGEN, H. C. V. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, *12*(24), 2305-2314. https://doi.org/10.1002/sim.4780122407. 62

MEINSHAUSEN, N. & BÜHHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*(3), 1436-1462. 63.

ZHAO, P. & YU, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, *7*(Nov), 2541-2563. http://www.jmlr.org/papers/v7/zhao06a.html. 63

TERNÈS, N., ROTOLO, F. & MICHIELS, S. (2016). Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. *Statistics in medicine*, *35*(15), 2561-2573. https://doi.org/10.1002/sim.6927. 63

TIBSHIRANI, R. (2011). Regression shrinkage and selection via the lasso : a retrospective. *Journal of the Royal Statistical Society : Series B (Statistical*

Methodology), 73(3), 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x. 64

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association, 101*(476), 1418-1429. https://doi.org/10.1198/016214506000000735. 65

Zhang, H. H. & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691-703. https://doi.org/10.1093/biomet/asm037. 65, 66

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*(456), 1348-1360. https://doi.org/10.1198/016214501753382273. 66

Michiels, S., Potthoff, R. F. & George, S. L. (2011). Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Statistics in medicine, 30*(13), 1502-1518. https://doi.org/10.1002/sim.4022. 72

Breheny, P. & Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface, 2*(3), 369-380. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904563/. 74

Huang, J., Breheny, P. & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science : a review journal of the Institute of Mathematical Statistics, 27*(4). https://doi.org/10.1214/12-STS392. 74

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics, 38*(2), 894-942. https://doi.org/10.1214/09-AOS729. 74

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics, 71*(3), 731-740. https://doi.org/10.1111/biom.12300. 75, 84

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *68*(1), 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x. 76

SIMON, N., FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231-245. https://doi.org/10.1080/10618600.2012.681250. 76

BOULESTEIX, A.-L., DE BIN, R., JIANG, X. & FUCHS, M. (2017). IPF-LASSO : Integrative-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and mathematical methods in medicine*, *2017*. https://doi.org/10.1155/2017/7691937. 76

PANG, H., TONG, T. & ZHAO, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, *65*(4), 1021-1029. https://doi.org/10.1111/j.1541-0420.2009.01200.x. 78

GENOVESE, C. & WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *64*(3), 499-517. https://doi.org/10.1111/1467-9868.00347. 82

PAWITAN, Y., MICHIELS, S., KOSCIELNY, S., GUSNANTO, A. & PLONER, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, *21*(13), 3017-3024. https://doi.org/10.1093/bioinformatics/bti448. 82

SIMON, N., FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2011). Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1. https://doi.org/10.18637/jss.v039.i05. 84

Friedman, J., Hastie, T., Simon, N. & Tibshirani, R. (2018). *glmnet : Lasso and Elastic-Net Regularized Generalized Linear Models* [R-package version 2.0-16]. https://cran.r-project.org/web/packages/glmnet. 84

Patrick, B. (2020). *grpreg : Regularization Paths for Regression Models with Grouped Covariates* [R package version 3.3.0]. https://CRAN.R-project.org/package=grpreg. 84

Boulesteix, A.-L. & Fuchs, M. (2015). *ipflasso : Integrative Lasso with Penalty Factors* [R package version 0.1]. https://CRAN.R-project.org/package=ipflasso. 84

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2019). *SGL : Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization* [R package version 1.3]. https://CRAN.R-project.org/package=SGL. 84

Schafer, J., Opgen-Rhein, R., Zuber, V., Ahdesmaki, M., Silva, A. P. D. & Strimmer., K. (2017). *corpcor : Efficient Estimation of Covariance and (Partial) Correlation* [R package version 1.6.9]. https://CRAN.R-project.org/package=corpcor. 84

Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine*, *32*(30), 5381-5397. http://onlinelibrary.wiley.com/doi/10.1002/sim.5958/full. 84

Blanche, P. (2019). *timeROC : Time-Dependent ROC Curve and AUC for Censored Survival Data* [R package version 0.4]. https://CRAN.R-project.org/package=timeROC. 84

Ignatiadis, M., Singhal, S. K., Desmedt, C., Haibe-Kains, B., Criscitiello, C., Andre, F., Loi, S., Piccart, M., Michiels, S. & Sotiriou, C. (2012). Gene modules and response to neoadjuvant chemotherapy in breast can-

cer subtypes : a pooled analysis. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology, 30*(16), 1996-2004. https://doi.org/10.1200/JCO.2011.39.5624. 96

Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data : does a single gene beat them all? *Bioinformatics, 24*(19), 2200-2208. https://doi.org/10.1093/bioinformatics/btn374. 98

Tang, Z., Lei, S., Zhang, X., Yi, Z., Guo, B., Chen, J. Y., Shen, Y. & Yi, N. (2019). Gsslasso Cox : a Bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information. *BMC Bioinformatics, 20*(1), 94. https://doi.org/10.1186/s12859-019-2656-1. 103

Zhang, L., Morris, J. S., Zhang, J., Orlowski, R. Z. & Baladandayuthapani, V. (2014). Bayesian Joint Selection of Genes and Pathways : Applications in Multiple Myeloma Genomics. *Cancer informatics, 13*, CIN-S13787. https://doi.org/10.4137/CIN.S13787. 103

Binder, H. & Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics, 10*(1), 18. https://doi.org/10.1186/1471-2105-10-18. 103

Liquet, B., De Micheaux, P. L., Hejblum, B. P. & Thiébaut, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics, 32*(1), 35-42. https://doi.org/10.1093/bioinformatics/btv535. 103

Obozinski, G., Jacob, L. & Vert, J.-P. (2011). Group lasso with overlaps : the latent group lasso approach. *arXiv preprint arXiv:1110.0413*. 103.

Chapitre 4

Favoriser la contrainte hiérarchique des interactions biomarqueurs-traitement

Sommaire

4.1 Introduction 113

4.2 Modèle avec interactions sous la contrainte hiérarchique . . 114

4.3 Méthode proposée 116

4.3.1 Lasso adaptatif 117

4.3.1.1 Choix de la pondération proposée 117

4.3.1.2 Étapes d'estimation de la pondération proposée 118

4.4 Méthodes alternatives 121

4.4.1 Lasso adaptatif avec la pénalisation ridge 121

4.4.2 Composite Minimax Concave Penalty (cMCP) 123

4.4.3 Group exponential lasso (gel) 123

4.4.4 Sparse Group Lasso (SGL) 123

4.5 Étude de simulation 124

4.5.1 Simulation de données 125

4.5.2 Choix des scénarios 125

4.5.3 Critères d'évaluation 127

4.5.4 Implémentation 129

4.5.5 Résultats 129

4.6 Application : cancer du sein précoce 139

4.7 Conclusion 143

4.8 Références 146

4.1 Introduction

La médecine de précision, également appelée médecine stratifiée, suppose que l'effet du traitement est hétérogène selon les caractéristiques des patients. Elle cherche à identifier les biomarqueurs ou les signatures génétiques distinguant les patients qui bénéficieront le plus d'un traitement. Cette hétérogénéité correspond à l'existence d'interactions entre l'effet du traitement et les biomarqueurs. En termes statistiques, la non comptabilisation de ces biomarqueurs prédictifs ou modificateurs de l'effet du traitement peut conduire à des erreurs dans la conception et l'analyse des essais cliniques en raison d'une mauvaise spécification du modèle de régression (Betensky et al. [2002], Buyse et Michiels [2013]).

Les cellules tumorales du cancer du sein hormonosensibles expriment dans la majorité des cas les récepteurs à l'oestrogène ($ER+$) et/ou à la progestérone ($PR+$). En effet, l'oestrogène et la progestérone sont deux hormones qui favorisent le développement des cellules mammaires cancéreuses. Ces récepteurs hormonaux ($ER+$ et $PR+$) sont les premiers biomarqueurs utilisés comme indicateurs pronostiques et prédictifs importants de la réponse à l'hormonothérapie. En particulier, l'expression d'au moins un des deux récepteurs permet de sélectionner les femmes qui bénéficieront du traitement (Stendahl et al. [2006], Delozier [2010]). Depuis des décennies, le tamoxifène fut la référence en matière d'hormonothérapie adjuvante pour le cancer du sein précoce à récepteurs hormonaux positifs ($HR+$). Il bloque les récepteurs des oestrogènes présents sur les cellules cancéreuses ainsi il inhibe la croissance de ces cellules. D'ailleurs, il a été démontré que le traitement pendant 5 ans par le tamoxifène adjuvant est associé à une diminution de 39% en moyenne du taux de récidive du cancer du sein sur 15 ans ((EBCTCG) [2011]).

Les interactions entre les biomarqueurs et le traitement peuvent être modélisées en utilisant un modèle de survie avec une structure hiérarchique. Dans la suite du chapitre, nous présentons dans un premier temps le modèle d'interaction hiérarchique biomarqueurs-traitement, ensuite les approches statistiques permettant de favoriser cette contrainte hiérarchique lors de l'identification des biomarqueurs prédictifs de l'effet du traitement. Puis, une étude de simulation comparant ces méthodes et enfin, une application de ces méthodes sur des données réelles du cancer du sein précoce.

4.2 Modèle avec interactions sous la contrainte hiérarchique

Soient X une matrice $n \times p$ des biomarqueurs standardisés et T une variable aléatoire représentant le traitement qui prend deux valeurs : $-1/2$ pour le bras contrôle et $+1/2$ pour le bras expérimental. Notons par $X_j T$, $(j = 1, \ldots, p)$, le produit élément par élément (*element-wise product*) entre le j^{e} biomarqueur et le traitement. Le modèle de Cox d'interaction hiérarchique biomarqueurs-traitement est défini comme suit :

$$h(t, T, X) = h_0(t) \exp\left(\alpha T + \sum_{j=1}^{p} \beta_j X_j + \sum_{j=1}^{p} \gamma_j X_j T\right), \tag{4.1}$$

où $\alpha, \beta = (\beta_1, \ldots, \beta_p)^T$ et $\gamma = (\gamma_1, \ldots, \gamma_p)^T$ représentent, respectivement, les coefficients de régression de l'effet du traitement, les effets pronostiques et les effets prédictifs des biomarqueurs. En effet, le premier terme de la partie exponentielle du modèle 4.1 correspond à l'effet propre du traitement, le second

terme correspond aux effets propres des biomarqueurs (*main effects*) et le troisième terme correspond aux interactions biomarqueurs-traitement.
La contrainte hiérarchique (Chipman [1996], Hamada et Wu [1992]) (*heredity* ou *hierarchically well-formulated*) appliquée au modèle 4.1 stipule qu'une interaction biomarqueur-traitement n'est incluse dans le modèle que si l'effet propre du biomarqueur correspondant est aussi dans le modèle. Cette contrainte se traduit mathématiquement par :

$$Si \quad \widehat{\gamma}_j \neq 0 \quad \Longrightarrow \quad \hat{\beta}_j \neq 0. \tag{4.2}$$

Plusieurs statisticiens comme Cox [1984], McCullagh [2019], et Bien et al. [2013] affirment que les modèles violant la contrainte hiérarchique sont inadéquats (non pertinents). Ainsi, l'objectif de ce travail de recherche est de favoriser la contrainte de la hiérarchie entre les effets propres et les interactions biomarqueurs-traitement dans la procédure de sélection des biomarqueurs prédictifs.
Les coefficients de régression du modèle semi-paramétrique de Cox (Cox [1972]) sont souvent estimés en maximisant la fonction de log-vraisemblance partielle $\ell(\beta, X)$. Toutefois, dans le cadre de données génomiques de grande dimension ($p \gg n$), les contraintes méthodologiques sont importantes si l'on veut respecter la hiérarchie. Notamment, le modèle 4.1 devient non identifiable c'est-à-dire qu'il n'existe pas une solution unique au problème d'estimation des paramètres de régression. Comme dit précédemment dans le chapitre 3, des méthodes de sélection de variables souvent utilisées dans le cadre des données en grande dimension sont les méthodes de régression pénalisée.

4.3 Méthode proposée

L'approche proposée est la suivante :

1. Créer des groupes (binômes) composés de l'effet propre du biomarqueur et de son interaction avec l'effet du traitement. En d'autres termes, créer p groupes, chaque groupe j étant composé de deux éléments $V_j = X_j$ et $V_{p+j} = X_j T$, $(j = 1, \ldots, p)$. Par conséquent, le modèle 4.1 s'écrit comme suit :

$$h(t, T, X) = h_0(t) \exp\left(\alpha T + \sum_{j=1}^{p} \left(\delta_j V_j + \delta_{p+j} V_{p+j}\right)\right), \tag{4.3}$$

   où δ_j est le j^{e} élément du vecteur δ des $2p$ coefficients de régression (p pour les effets propres et p pour les interactions avec le traitement) et V_j est la j^{e} colonne de la matrice

$$\boldsymbol{V} = (\boldsymbol{X} \mid \boldsymbol{X}T),$$

   qui est une matrice $n \times p'$, avec les colonnes 1 à p qui sont les effets propres des biomarqueurs (matrice $\boldsymbol{X}$) et les colonnes $p+1$ à $2p$ qui sont leurs interactions avec les traitement (matrice $\boldsymbol{XT}$).

2. Faire la sélection à la fois au niveau des groupes et à l'intérieur des groupes tout en favorisant la contrainte de la hiérarchie. Nous avons utilisé la méthode de régression pénalisée lasso adaptatif (Zou [2006], H. H. Zhang et Lu [2007]), et nous avons proposé des poids inversement proportionnels à la force de l'interaction entre le traitement et le biomarqueur.

Le choix de cette méthode est aussi justifié par le fait que les tailles de l'effet propre du biomarqueur et de l'interaction peuvent être très différentes et la méthode lasso adaptatif permet de les pénaliser différemment.

4.3.1 Lasso adaptatif

La méthode lasso adaptatif, comme décrite dans le chapitre 3 (section 3.3), est une extension de la méthode lasso standard (TIBSHIRANI [1996]). Elle introduit des poids adaptatifs W à la fonction de pénalisation de la log-vraisemblance partielle pour pénaliser chaque coefficient de régression δ différemment. La fonction de log-vraisemblance pénalisée du lasso adaptatif s'écrit comme :

$$\ell_p(\delta, T, X) = \ell(\delta, T, X) - \lambda \left(\sum_{j=1}^{p} \left(W_j \left|\delta_j\right| + W_{p+j} \left|\delta_{p+j}\right| \right) \right), \tag{4.4}$$

où λ est le paramètre de régularisation calculé par une validation croisée à k blocs. Pour chaque groupe j, $W^{(j)} = \left(W_j, W_{p+j}\right)$ représentent les poids adaptatifs spécifiques pour l'effet propre du biomarqueur et son interaction avec le traitement, estimés dans une étape préliminaire.

4.3.1.1 Choix de la pondération proposée

Nous avons proposé des poids adaptatifs basés sur la statistique du test du rapport de vraisemblance (*likelihood ratio test*, LRT). Le test du rapport de vraisemblance, plus général que le test de Wald (section 3.3.2), permet de tester la significativité d'un vecteur de paramètres. En particulier, il peut être utilisé pour évaluer la qualité de l'ajustement de deux modèles statistiques emboîtés [1] : M_0 (modèle de référence) et M_1 (modèle candidat) avec M_0 est emboîté dans M_1. Les deux hypothèses nulle H_0 et alternative H_1 du test du rapport de

1. Un modèle est emboîté dans un autre modèle quand il est sous-ensemble de ce dernier.

vraisemblance s'écrivent, respectivement, comme suit :

H_0 : aucune différence entre les deux modèles M_0 et M_1

H_1 : le modèle candidat M_1 est meilleur

Sous l'hypothèse nulle, la statistique du test du rapport de vraisemblance comparant les modèles M_0 et M_1 suit une distribution Chi-deux à q degrés de liberté (nombre de paramètres additionnels du M_1) et s'écrit comme :

$$\Lambda_{M_1-M_0} = -2 \times ln\left(\frac{L(M_0)}{L(M_1)}\right) = 2 \times (\ell(M_1) - \ell(M_0)) \sim \chi^2(q).$$

$L(.)$ et $\ell(.)$ représentent les fonctions de vraisemblance et de log-vraisemblance, respectivement. Le nombre de degrés de libertés q de la statistique du Chi-deux est obtenu en faisant la différence entre le nombre de paramètres à estimer sous H_1 moins le nombre de paramètres à estimer sous H_0. La valeur de la fonction de vraisemblance du modèle M_1 ($\ell(M_1)$) est supérieure ou égale à celle du modèle M_0 ($\ell(M_0)$) car le modèle M_0 est emboîté dans le modèle M_1. Par conséquent, la statistique du test du rapport de vraisemblance $\Lambda_{M_1-M_0}$ est positive. Pour des grandes valeurs de $\Lambda_{M_1-M_0}$, on favorise l'hypothèse H_1 par rapport à H_0.

4.3.1.2 Étapes d'estimation de la pondération proposée

Pour estimer les poids adaptatifs $W^{(j)}$, $(j = 1, \ldots, p)$, nous avons considéré les modèles de Cox suivants :

— Le modèle de référence (avec seulement l'effet du traitement)

$M_0 : h(t,T) = h_0(t) \exp(\alpha T)$.

— Le modèle avec l'effet propre du biomarqueur

$M_1 : h(t,T,X) = h_0(t) \exp\left(\alpha T + \beta_j X_j\right), \quad j = 1, \ldots, p.$

— Le modèle avec l'effet propre du biomarqueur et l'interaction biomarqueur-traitement

$$M_2 : h(t,T,X) = h_0(t)\exp\left(\alpha T + \beta_j X_j + \gamma_j X_j T\right), \quad j = 1,\ldots,p.$$

Le modèle M_0 est emboîté dans les modèles M_1 et M_2 ; le modèle M_1 est emboîté dans M_2.

A partir de ces modèles, nous avons considéré les hypothèses suivantes :

— Pour les modèles M_0 $versus$ M_2

$$H_0 : \beta_j = \gamma_j = 0,$$
$$H_1 : \left\{\beta_j \text{ et/ou } \gamma_j\right\} \neq 0,$$

— Pour les modèles M_1 $versus$ M_2

$$H_0 : \gamma_j = 0,$$
$$H_1 : \gamma_j \neq 0,$$

Pour chaque biomarqueur, nous avons calculé les statistiques du test du rapport de vraisemblance partielle entre les modèles M_0 et M_2 ($\Lambda_{M_2-M_0}$) et entre M_1 et M_2 ($\Lambda_{M_2-M_1}$).

Pour chaque groupe j, ($j = 1,\ldots,p$), les poids adaptatifs accordés à l'effet propre du biomarqueur W_j et à l'interaction biomarqueur-traitement W_{p+j} ont été défini comme suit :

$$W^{(j)} = \left(W_j, W_{p+j}\right) = \left(\frac{1}{\Lambda^{(j)}_{M_2-M_0}}, \frac{1}{\Lambda^{(j)}_{M_2-M_1}}\right). \tag{4.5}$$

Plus la statistique du test de vraisemblance partielle (Λ) est élevée plus la pondération ($W^{(j)}$) est faible. Par exemple, si la valeur de $\Lambda_{M_2-M_0}$ est élevée cela signifie qu'au moins l'un des deux (l'effet propre du biomarqueur et l'interaction biomarqueur-traitement) semblerait être non nul ($\beta_j \neq 0$ et/ou $\gamma_j \neq 0$). Par

conséquent, la pondération accordée au biomarqueur dans le modèle pénalisé ($W_j = 1/\Lambda^{(j)}_{M_2-M_0}$) est faible, ce qui augmente les chances qu'il soit sélectionné dans le modèle final quelle que soit la valeur de l'effet propre de ce biomarqueur.

En d'autres termes, ce choix de pondérations proposées permet de favoriser la contrainte de la hiérarchie des interactions biomarqueurs-traitement. En effet, considérons les cas de figure suivants :

— Si un biomarqueur j est un biomarqueur prédictif, c'est-à-dire l'interaction biomarqueur-traitement est non nulle ($\gamma_j \neq 0$) alors les valeurs des statistiques du test de vraisemblance partielle ($\Lambda_{M_2-M_0}$ et $\Lambda_{M_2-M_1}$) sont élevées. En conséquence, les valeurs des poids accordés à l'effet propre du biomarqueur et à l'interaction (W_j et W_{p+j}) sont faibles. Cela favorise la sélection de l'interaction biomarqueur-traitement et du biomarqueur correspondant, indépendamment de la valeur de l'effet propre de ce biomarqueur.

— D'autre part, si un biomarqueur j est un biomarqueur pronostique, c'est-à-dire a un effet propre important mais n'interagit pas avec le traitement ($\beta_j \neq 0$ et $\gamma_j = 0$) alors la valeur de $\Lambda_{M_2-M_0}$ est élevée mais $\Lambda_{M_2-M_1}$ est faible. Donc, la valeur du poids W_j est faible et du poids W_{p+j} est élevée, ce qui favorise la sélection du biomarqueur pronostique uniquement sans l'interaction.

Dans la suite du manuscrit, la méthode lasso adaptatif avec les pondérations proposées basées sur la statistique du test de vraisemblance partielle est nommée lasso adaptatif (LRT).

4.4 Méthodes alternatives

Nous avons comparé la méthode proposée avec les quatre méthodes alternatives suivantes :

4.4.1 Lasso adaptatif avec la pénalisation ridge

Dans le contexte d'identification des interactions, Ternes, Rotolo, Heinze et al. [2017] ont comparé plusieurs approches pour la sélection de biomarqueurs prédictifs. Parmi ces approches, ils ont étudié des extensions du lasso standard qui pénalisent à la fois les effets spécifiques des biomarqueurs et les interactions biomarqueur-traitement avec des extensions du lasso standard. Ils ont montré que la méthode lasso adaptatif avec des pondérations estimées par la méthode ridge (lasso adaptatif (ridge)), (présentée ci-dessous) a des bonnes performances en termes de sélection et prédiction.

Pénalisation ridge

La régression ridge est une des méthodes de pénalisation introduite par Hoerl et Kennard [1970]. Elle est souvent utilisée dans le cas de forte corrélation entre des variables. La méthode ridge introduit un terme de pénalité à la fonction de log-vraisemblance partielle du modèle 4.3 pour limiter l'instabilité des coefficients. En effet, le terme de pénalité de ridge correspond à la norme L_2 des coefficients de régression et s'écrit comme :

$$p(\lambda, \delta) = \lambda \|\delta\|_2^2 = \lambda \left(\sum_{j=1}^{p} \left(\delta_j^2 + \delta_{p+j}^2 \right) \right). \tag{4.6}$$

La fonction de log-vraisemblance partielle pénalisée de la méthode ridge est

définie de la manière suivante :

$$\ell_p(\delta, T, X) = \ell(\delta, T, X) - \lambda \left(\sum_{j=1}^{p} \left(\delta_j^2 + \delta_{p+j}^2 \right) \right). \tag{4.7}$$

où λ est le paramètre de régularisation qui permet de contrôler l'impact de la pénalité ($\lambda \geq 0$). En augmentant la valeur du paramètre λ, les valeurs des coefficients de régression tendent vers zéro. Contrairement à la méthode lasso standard, la méthode ridge n'annule aucun coefficient de régression. Cela rend son utilisation limitée en ce qui concerne la sélection de variables en grande dimension. En revanche, la pénalisation ridge permet de réduire la complexité du modèle en réduisant la variance. Elle donne aussi de bonnes performances prédictives (Hoerl et Kennard [1970]).

La maximisation de la fonction de log-vraisemblance partielle pénalisée 4.7 permet d'estimer coefficients de régression $\tilde{\delta}_j^R$ et $\tilde{\delta}_{p+j}^R$. Ensuite, les poids adaptatifs accordés à l'effet propre du biomarqueur et à l'interaction biomarqueur-traitement sont définis comme l'inverse de la valeur absolue de ces coefficients de régression :

$$W^{(j)} = \left(W_j, W_{p+j}\right) = \left(\frac{1}{\tilde{\delta}_j^R}, \frac{1}{\tilde{\delta}_{p+j}^R} \right). \tag{4.8}$$

Ainsi, la fonction de log-vraisemblance pénalisée du lasso adaptatif (ridge) s'écrit comme suit :

$$\ell_p(\delta, T, X) = \ell(\delta, T, X) - \lambda \left(\sum_{j=1}^{p} \left(\frac{1}{\tilde{\delta}_j^R} \left|\delta_j\right| + \frac{1}{\tilde{\delta}_{p+j}^R} \left|\delta_{p+j}\right| \right) \right). \tag{4.9}$$

Les autres méthodes alternatives présentées dans la suite sont détaillées dans le chapitre précédent 3 (section 3.4).

4.4.2 Composite Minimax Concave Penalty (cMCP)

La méthode (*composite Minimax Concave Penalty*) ou cMCP (Breheny et Huang [2009], Huang et al. [2012]) effectue la sélection de variables à deux niveaux : à la fois des groupes et au sein des groupes sélectionnés. En considérant le modèle 4.3 la fonction de log-vraisemblance pénalisée du cMCP s'écrit comme :

$$\ell_p(\delta, T, X) = \ell(\delta, T, X) - \sum_{j=1}^{p} f_O\left(f_I\left(\delta_j\right) + f_I\left(\delta_{p+j}\right)\right). \tag{4.10}$$

où f_O et f_I sont les fonctions de pénalité MCP (C.-H. Zhang et al. [2010]) extérieur (*outer*) et intérieur (*inner*), respectivement. Comme dit précédemment, la pénalité MCP est définie sur $[0, \infty)$ comme suit :

$$f_{\lambda,a}(\delta) = \begin{cases} \lambda\delta - \frac{\delta^2}{2a}, & \text{si } \delta \leq a\lambda \\ \frac{1}{2}a\lambda^2, & \text{si } \delta > a\lambda \end{cases}, \text{ avec } \lambda \geq 0. \tag{4.11}$$

4.4.3 Group exponential lasso (gel)

La méthode (*group exponential lasso*) ou gel (Breheny [2015]) permet également de faire la sélection à deux niveaux. La fonction de log-vraisemblance pénalisée de la méthode gel est similaire à celle du cMCP, avec une pénalité extérieure f_O égale à la pénalité exponentielle et une pénalité intérieure f_I égale à la pénalité du lasso standard.

4.4.4 Sparse Group Lasso (SGL)

La méthode (*Sparse Group Lasso*) ou SGL (Simon et al. [2013]) favorise aussi la sélection de variables à deux niveaux (*groupwise sparsity* et *within group*

sparsity). Sa fonction de log-vraisemblance pénalisée est définie comme suit :

$$\begin{aligned}\ell_p(\delta, T, X) &= \ell(\delta, T, X) - (1-\alpha)\lambda \sum_{j=1}^{p} \sqrt{2} \left\| \delta^{(j)} \right\|_2 - \alpha\lambda \|\delta\|_1 \\ &= \ell(\delta, T, X) - (1-\alpha)\lambda \sum_{j=1}^{p} \sqrt{2\left(\delta_j + \delta_{p+j}\right)} - \alpha\lambda \sum_{j=1}^{p} \left|\left(\delta_j + \delta_{p+j}\right)\right|.\end{aligned} \tag{4.12}$$

Le paramètre α est compris dans l'intervalle [0,1]. Pour les valeurs 1 et 0, la méthode SGL se résume soit à la méthode lasso standard ou soit au lasso groupé (*group lasso*) (Yuan et Lin [2006]), respectivement.
La méthode lasso groupé peut imposer la contrainte de la hiérarchie, mais au prix de la sélection ou de l'élimination en bloc des paires (effet propre du biomarqueur et interaction biomarqueurs-traitement). Cependant, cette méthode sélectionne de nombreuses fausses interactions pour les biomarqueurs pronostiques.

4.5 Étude de simulation

Afin d'évaluer la capacité de sélection et de prédiction des méthodes présentées précédemment, une étude de simulation a été réalisée avec six scénarios. Nous avons étudié la capacité des méthodes à respecter la contrainte hiérarchique des interactions biomarqueurs-traitement c'est-à-dire à détecter les interactions avec les effets propres des biomarqueurs correspondants. Nous avons également évalué si les gènes identifiés peuvent suffisamment discriminer le bénéfice du traitement chez les futurs patients.

4.5.1 Simulation de données

Les données ont été générées d'une manière similaire à celle présentée dans le chapitre 3 (section 3.5). Chaque jeu de données généré contient $n = 1500$ patients et $p = 500$ biomarqueurs (expression de gènes). Les patients ont été assignés aléatoirement à chaque bras de traitement avec une probabilité de $1/2$. Le traitement a été codé $+1/2$ pour le bras expérimental et $-1/2$ pour le bras contrôle. Les biomarqueurs ont été générés à partir d'une distribution gaussienne centrée et réduite (les moyennes $\mu_1 = \cdots = \mu_p = 0$ et les écarts types $\sigma_1 = \cdots = \sigma_p = 1$). La structure de corrélation des biomarqueurs a été définie comme autorégressive de blocs de 20 biomarqueurs, dans chaque bloc la corrélation entre deux biomarqueurs i et j a été fixée à $\rho_{ij} = 0,7^{|i-j|}$. Nous avons généré le temps de survie à l'aide d'une distribution exponentielle avec une survie médiane d'un an. Les temps de censure ont été générés à partir d'une distribution uniforme U(2,5), reflétant un essai avec une période de recrutement de trois ans et un temps de suivi de deux ans.

4.5.2 Choix des scénarios

Différents scénarios ont été générés en variant q_{Po} le nombre de biomarqueurs pronostiques et q_{Pe} le nombre de biomarqueurs prédictifs (modificateur de l'effet du traitement). Nous avons considéré un scénario nul ($q_{Pe} = 0$) dans lequel aucun biomarqueur n'interagit avec le traitement :

— Scénario 1 avec aucun signal ($q_{Po} = 0$ et $q_{Pe} = 0$).

Les trois scénarios alternatifs avec au moins un biomarqueur qui interagit avec le traitement sont les suivants :

Les scénarios 2 et 3 ne contiennent que des biomarqueurs prédictifs.

— Scénario 2 avec $q_{Pe} = 1$ biomarqueur prédictif.

— Scénario 3 avec $q_{Pe} = 10$ biomarqueurs prédictifs.

— Scénario 4, plus réaliste, contient à la fois $q_{Po} = 10$ biomarqueurs pronostiques et $q_{Pe} = 10$ autres biomarqueurs prédictifs différents.

Les biomarqueurs pronostiques et prédictifs ont été générés avec un effet de $\beta = \gamma = ln(0,5)$. Pour chaque scénario, 500 réplications ont été effectuées. Le tableau 4.1 résume les six scénarios de l'étude de simulation.

TABLEAU 4.1 : Scénarios de l'étude de simulation

	Scénarios	$HR = \exp(\beta)$		Médiane de survie (années)		Taux de censure moyen	
		T^-	T^+	T^-	T^+	T^-	T^+
1	Aucun effet	1	1	1	1	10%	10%
2	$q_{Pe} = 1$ biomarqueur prédictif	1	0,5	1	1	12%	12%
3	$q_{Pe} = 10$ biomarqueurs prédictifs	1	0,5	1	1	20%	20%
4	$q_{Pe} = 10$ biomarqueurs prédictifs + $q_{Po} = 10$ biomarqueurs pronostiques	1 0,5	0,5	1	1	30%	31%

HR (Hazard Ratio) : rapport de risque instantané, T^- : bras contrôle, T^+ : bras expérimental, q_{Po} : nombre de biomarqueurs pronostiques, q_{Pe} : nombre de biomarqueurs prédictifs

Nous rappelons que le modèle proposé à la section 4.2 (équation 4.1) est défini avec la matrice $\boldsymbol{V} = (\boldsymbol{X} \mid \boldsymbol{X}T)$. A partir des 500 biomarqueurs simulés, nous avons calculé le produit élément par élément entre chaque colonne de la matrice X des biomarqueurs et le traitement. Par la suite, nous avons défini la matrice V regroupant la matrice X et XT. Ainsi, nous avons considéré 1000 variables pour chaque jeu de données.

4.5.3 Critères d'évaluation

Dans cette étude de simulation, l'objectif principal était d'évaluer la capacité des différentes méthodes à sélectionner les vrais interactions (biomarqueurs prédictifs) avec les effets propres des biomarqueurs correspondants (respecter la contrainte hiérarchique des interactions biomarqueurs-traitement). Comme dans le chapitre 3 (tableau 3.1), nos critères d'évaluation étaient les suivants :

— Pour le scénario nul 1 ($q_{Po} = 0$ et $q_{Pe} = 0$), nous avons calculé le nombre de faux positifs (FP) et le taux de fausses découvertes ($FDR = FP/(FP + VP)$) des effets propres et des interactions. En effet, le FDR est égale à 0 si aucun biomarqueur n'est sélectionné ou 1 si au moins un effet propre (ou une interaction) est sélectionné(e), respectivement.

— Pour les scénarios alternatifs 2, 3, et 4 ($q_{Pe} \neq 0$), nous avons calculé le nombre des interactions sélectionnées, le nombre des effets propres sélectionnés correspondant à des interactions sélectionnées, le nombre des vrais positifs (VP), FP, le FDR et le taux de faux négatifs ($FNR = FN/(FN + VP)$) des interactions et des effets propres correspondant.

Selon ces critères, les méthodes qui ont une bonne capacité de sélection sont celles qui ont à la fois la meilleure balance FDR-FNR des interactions et des effets propres correspondants.

Nous avons évalué également la capacité prédictive de toutes les méthodes. van Klaveren et al. [2018] ont proposé la concordance entre le bénéfice du traitement prédit et observé (*c-for-benefit*) comme une mesure de discrimination pour évaluer la capacité du modèle à distinguer les patients qui bénéficient davantage du traitement de ceux qui en bénéficient moins. En effet, les bénéfices du traitement chez un patient individuel sont inobservables puisque le patient est affecté à l'un des deux bras de traitement et seul un des deux ré-

sultats contrefactuels est observé. Pour chaque patient, nous avons calculé le bénéfice du traitement prédit comme la différence de risque entre les deux bras de traitement.

Le bénéfice du traitement observé a été défini comme la différence de résultats entre deux patients appariés selon leurs bénéfices prédits mais avec des affectations de traitement différentes (bras contrôle et bras expérimental). En effet, pour chaque paire de patients (i appartenant au bras contrôle, j appartenant au bras expérimental) appariés selon leurs bénéfices prédits, nous avons calculé le bénéfice du traitement observé avec la formule suivante :

$$\mathbb{1}_{Statut_i} \times \mathbb{1}_{\{S_i<S_j\}} - \mathbb{1}_{Statut_j} \times \mathbb{1}_{\{S_j<S_i\}} \tag{4.13}$$

avec $\mathbb{1}_{Statut}$ est l'indicatrice du statut du patient qui vaut 1 si le patient est décédé et 0 sinon et $\mathbb{1}_{\{S_i<S_j\}}$ est l'indicatrice de survie du patient i, elle prend deux valeurs : 1 si le temps de survie patient i (S_i) est inférieur au temps de survie patient j (S_j) et 0 sinon. De ce fait, le premier terme de l'équation $\mathbb{1}_{Statut_i} \times \mathbb{1}_{\{S_i<S_j\}}$ vaut 1 si le patient i du bras contrôle est décédé avant le patient j du bras expérimental et 0 sinon. Le raisonnement est le même pour le second terme de l'équation (pour le patient j).

Pour chaque paire de patient, le bénéfice du traitement expérimental observé peut prendre l'une des valeurs suivantes :

- 1 : effet bénéfique
- 0 : aucun effet
- -1 : effet délétère

La statistique de concordance pour le bénéfice (C-pour-bénéfice) correspond à la concordance en termes de bénéfice du traitement entre toutes les paires possibles de patients qui ont reçu des traitements différents mais dont le bénéfice du traitement prédit est similaire. Par exemple, supposant que pour deux

paires de patients, la première paire a un bénéfice prédit moyen plus élevé que la deuxième paire de patient. Il existe une concordance entre les deux paires de patients si et seulement si la première paire a aussi un bénéfice observé plus élevé que la deuxième paire de patients. En d'autres termes, la statistique C-pour-bénéfice est définie comme la probabilité que parmi deux paires de patients, la paire présentant le plus grand bénéfice observé présente également un bénéfice prédit moyen plus élevé.
Les méthodes qui ont une bonne capacité de prédiction selon ce critère sont celles qui ont une valeur de C-pour-bénéfice élevée (proche de 1), c'est-à-dire celles qui discriminent plus les patients qui sont davantage susceptibles de bénéficier du traitement expérimental.

4.5.4 Implémentation

Les packages suivants du logiciel `R` ont été utilisés pour implémenter les méthodes présentées précédemment : `glmnet` (Simon et al. [2011] et Friedman et al. [2018]) pour la méthode lasso adaptatif (LRT) ; `biospear` (Ternes, Rotolo et Michiels [2017]) pour la méthode lasso adaptatif (ridge) ; `grpreg` (Patrick [2020]) pour les méthodes cMCP et gel avec $\tau = 1/3$ (Breheny [2015]) ; `SGL` (Noah et al. [2019]) pour la méthode SGL avec la valeur par défaut du paramètre $\alpha = 0,95$. Enfin, nous avons utilisé le code `R cbenefit` disponible en ligne sur la plate forme Github pour calculer la statistique C-pour-bénéfice dans le cadre de la survie.

4.5.5 Résultats

Les résultats de l'étude de simulation sont présentés dans les tableaux (4.2, 4.3, 4.4) et les figures (4.1, 4.2).

Évaluation de la sélection des interactions biomarqueurs-traitement

Le tableau 4.2 et la figure 4.1 résument les résultats du n_{Pe} nombre d'interaction biomarqueurs-traitement sélectionnées, les VP, les FP, le FDR, et le FNR des interactions.

Scénario nul 1 avec aucun signal, ($q_{Pe} = 0$ biomarqueur prédictif et $q_{Po} = 0$ biomarqueur pronostique).

La méthode lasso adaptatif (LRT) avait en moyenne le nombre le plus élevé d'interactions sélectionnées ($n_{Pe} = 9,96$) en comparant avec les autres méthodes. La méthode SGL n'a sélectionné qu'une seule interaction dans un jeu de données parmi les 500 réplications.

Globalement, à travers les scénarios alternatifs, toutes les méthodes ont identifié les interactions actives.

Scénario 2 avec une seule interaction ($q_{Pe} = 1$).

Les méthodes gel et SGL avaient une valeur de $VP = 1$ et les plus faibles valeurs de FP, cela se traduit par des faibles valeurs FDR/FNR. A noter qu'elle avait 31 itérations qui n'ont pas convergé. La méthode lasso adaptatif (LRT) avait, en moyenne, le nombre le plus élevé de FP en comparant avec les autres méthodes.

Scénario 3 avec un nombre plus élevé de biomarqueurs prédictifs ($q_{Pe} = 10$).

La méthode gel avait les plus faibles valeurs $FDR/FNR = 0,04$. Tandis que les méthodes lasso adaptatif (ridge) et SGL avaient les plus grandes valeurs de FP ainsi de FDR par rapport aux autres méthodes.

Scénario nul 4, plus réaliste, ($q_{Pe} = 10$ biomarqueurs prédictifs et $q_{Po} = 10$ biomarqueurs pronostiques).

Les méthodes lasso adaptatif (LRT) et cMCP avaient les meilleurs balances $FDR-FNR$ avec des valeurs égales à 0,32/0,00 et 0,29/0,00, respectivement. Les mé-

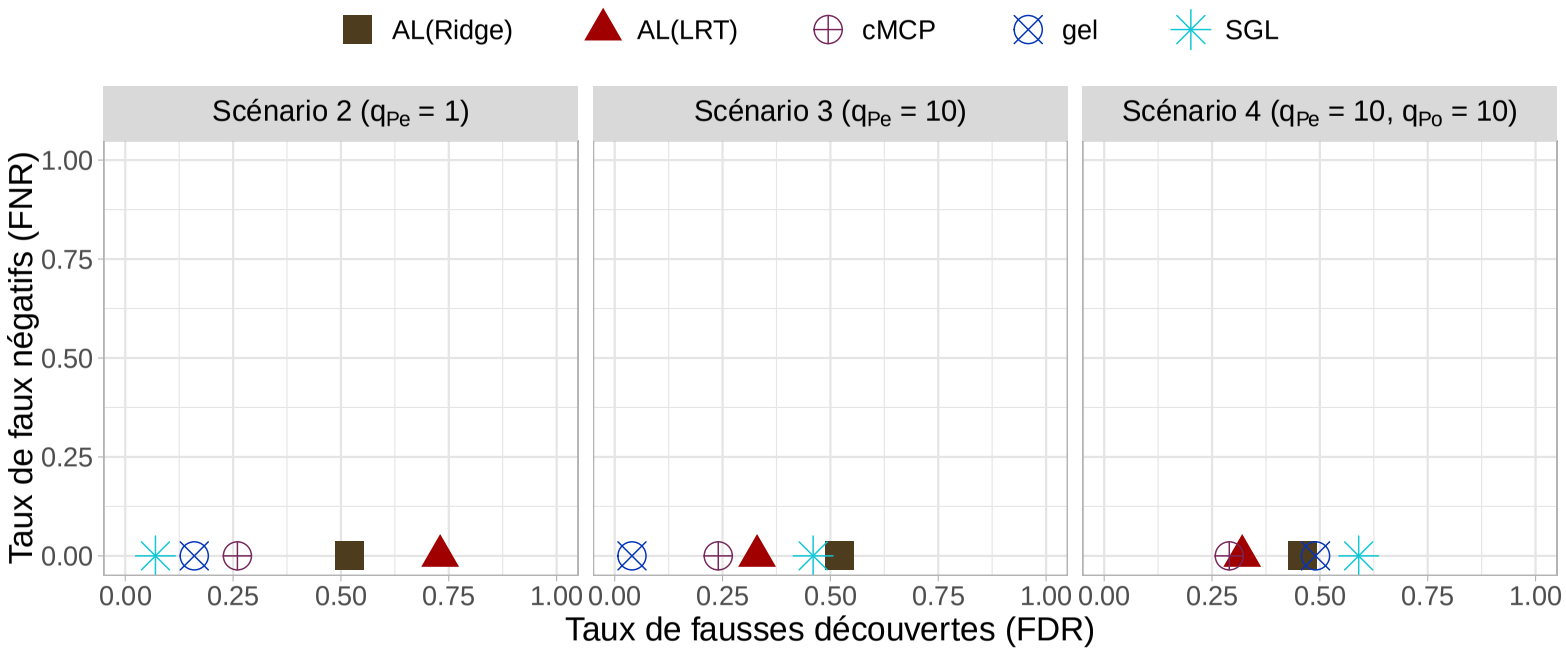
thodes gel et lasso adaptatif (ridge) avaient une performance similaire avec des valeurs FDR/FNR de 0,49/0,00 et 0,46/0,00 respectivement. La méthode SGL, avait une valeur de $FP = 25,32$ la plus élevée.

TABLEAU 4.2 : Capacité de sélection des interactions biomarqueurs-traitement dans tous les scénarios

		AL(ridge)	AL(LRT)	cMCP	gel	SGL
Scénario 1 Aucun effet	n_{Pe}	1,95	9,96	0,78	0,46	0
	FP / VP	1,95/-	9,96/-	0,78/-	0,46/-	0/-
	FDR / FNR	0,50/-	0,95/-	0,24/-	0,20/-	0/-
Scénario 2 q_{Pe} = 1 biomarqueur prédictif	n_{Pe}	4,34	9,32	2,28	1,58	1,17
	FP / VP	3,34/1,00	8,32/1,00	1,28/1,00	0,58/1,00	0,17/1,00
	FDR / FNR	0,52/0,00	0,73/0,00	0,26/0,00	0,16/0,00	0,07/0,00
Scénario 3 q_{Pe} = 10 biomarqueurs prédictifs	n_{Pe}	22,58	15,39	13,81	10,52	19,33
	FP / VP	12,58/10,00	5,39/10,00	3,81/10,00	0,52/10,00	9,33/10,00
	FDR / FNR	0,52/0,00	0,33/0,00	0,24/0,00	0,04/0,00	0,46/0,00
Scénario 4 q_{Pe} = 10 biomarqueurs prédictifs + q_{Po} = 10 biomarqueurs pronostiques	n_{Pe}	20,11	15,13	14,58	19,67	25,32
	FP / VP	10,11/10,00	5,15/9,98	4,58/10,00	9,69/9,99	15,32/10,00
	FDR / FNR	0,46/0,00	0,32/0,00	0,29/0,00	0,49/0,00	0,59/0,00

Résultats avec p = 500 biomarqueurs et n = 1500 patients. Quantités moyennes sur 500 réplications. q_{Pe} : nombre de biomarqueurs prédictifs, q_{Po} : nombre biomarqueurs pronostiques, n_{Pe} : nombre d'interactions sélectionnées, FP : faux positifs, VP : vrais positifs, FDR : taux de fausses découvertes, FNR : taux de faux négatifs, AL : lasso adaptatif, LRT : test du rapport de vraisemblance

FIGURE 4.1 : *Capacité de sélection des interactions biomarqueurs-traitement dans les scénarios alternatifs*



Résultats avec $p = 500$ biomarqueurs et $n = 1500$ patients. Quantités moyennes sur 500 réplications. q_{Pe} : nombre de biomarqueurs prédictifs, q_{Po} : nombre biomarqueurs pronostiques, FDR : taux de fausses découvertes, FNR : taux de faux négatifs, AL : lasso adaptatif, LRT : test du rapport de vraisemblance

Évaluation de la sélection des effets propres de biomarqueurs prédictifs

Le tableau 4.3 et la figure 4.2 résument les résultats du n_{Po} nombre d'effets propres sélectionnés correspondant à des interactions sélectionnées, les VP, les FP, le FDR, et le FNR des effets propres des biomarqueurs prédictifs sélectionnés.

Scénario nul 1 avec aucun signal, ($q_{Pe} = 0$ biomarqueur prédictif et $q_{Po} = 0$ biomarqueur pronostique).

La méthode lasso adaptatif (LRT) a sélectionné en moyenne $n_{Po} = 3,96$ effets propres correspondant à des interactions sélectionnées, pendant que les autres méthodes n'ont pas sélectionné presque aucun effet propre (n_{Po} varie entre 0 et 0,02).

Scénario 2 avec une seule interaction ($q_{Pe} = 1$).

La méthode lasso adaptatif (LRT) avait un nombre d'effets propres sélectionnés correspondant à des interactions sélectionnées le plus élevé $n_{Po} = 5,04$ avec un nombre de $FDR = 0,65$ le plus élevé et de $FNR = 0,12$ le plus faible, alors que les autres méthodes ne sélectionnaient pas les effets propres des interactions sélectionnées dans la plupart des cas avec des nombres moyens de FDR variant entre $0,01$ et $0,08$ et de FNR entre $0,65$ et 1.

Scénario 3 avec un nombre plus élevé de biomarqueurs prédictifs ($q_{Pe} = 10$).

Le lasso adaptatif (LRT) avait la valeur la plus élevée du nombre d'effets propres sélectionnés $n_{Po} = 9,43$, en comparant avec les autres méthodes. Pour cette méthode, le FDR était de $0,27$ et le FNR était de $0,33$. D'autre part, la méthode gel avait des valeurs de FDR/FNR égales à $0/0,43$ avec un nombre d'effets propres correspondant à des interactions sélectionnées $n_{Po} = 5,71$.

Scénario nul 4, plus réaliste, ($q_{Pe} = 10$ biomarqueurs prédictifs et $q_{Po} = 10$ biomarqueurs pronostiques).

La méthode lasso adaptatif (LRT) semble avoir la meilleure balance $FDR-FNR =$

0, 28/0, 43. Bien que la méthode gel avait une valeur de VP similaire à celle du lasso adaptatif (LRT) et la plus élevée, elle avait un nombre de $FP = 9,53$ le plus élevé en comparant des autres méthodes.
D'une manière générale, la méthode lasso adaptatif (LRT) permet de respecter le plus la contrainte hiérarchique des interactions biomarqueurs-traitement au prix d'une plus grande probabilité de sélectionner des faux biomarqueurs.

Évaluation de la capacité prédictive des méthodes

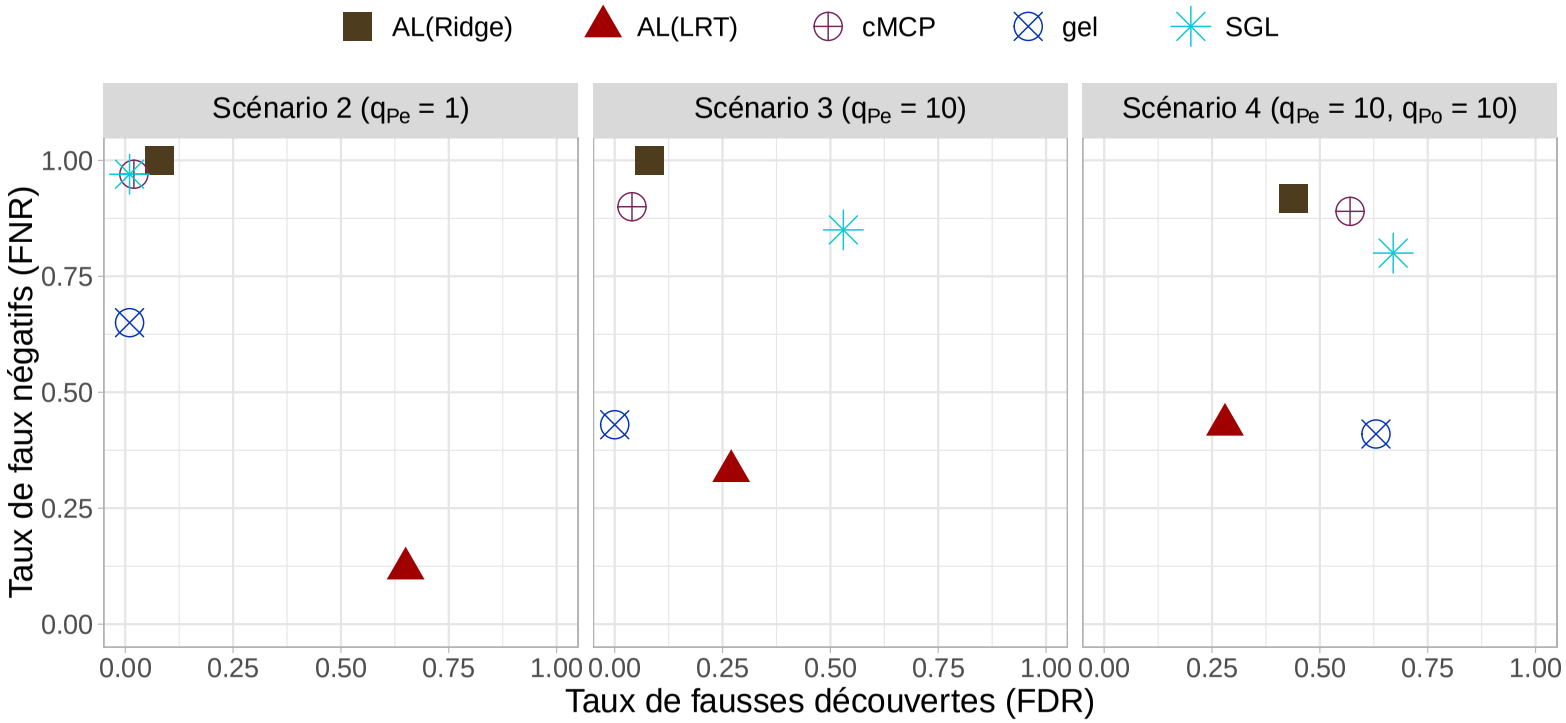
Le tableau 4.4 résume la valeur moyenne de la statistique de concordance pour bénéfice et son erreur type empirique. Dans l'ensemble, toutes les méthodes avaient des valeurs de C-pour-bénéfice similaires sauf la méthode SGL qui avait en moyenne une statistique C-pour-bénéfice autour de 0,5. Les méthodes lasso adaptatif (LRT) et cMCP avaient les valeurs les plus élevées dans tous les scénarios alternatifs, même si les différences entre méthodes étaient souvent très petites.

TABLEAU 4.3 : Capacité de sélection des effets propres des biomarqueurs prédictifs (interactions) dans tous les scénarios

		AL(ridge)	AL(LRT)	cMCP	gel	SGL
Scénario 1	n_{Po}	0	3,96	0,01	0,02	0
Aucun effet	FP / VP	0/-	3,96/-	0,01/-	0,02/-	0/-
	FDR / FNR	0/-	0,85/-	0,01/-	0,02/-	0/-
Scénario 2	n_{Po}	0,08	5,04	0,05	0,37	0,04
q_{Pe} = 1 biomarqueur prédictif	FP / VP	0,08/0,00	4,16/0,88	0,02/0,03	0,01/0,35	0,01/0,03
	FDR / FNR	0,08/1,00	0,65/0,12	0,02/0,97	0,01/0,65	0,01/0,97
Scénario 3	n_{Po}	0,08	9,43	1,07	5,71	3,69
q_{Pe} = 10 biomarqueurs prédictifs	FP / VP	0,08/0,00	2,76/6,67	0,07/1,00	0,01/5,70	2,14/1,54
	FDR / FNR	0,08/1,00	0,27/0,33	0,04/0,90	0,00/0,43	0,53/0,85
Scénario 4	n_{Po}	2,04	8,18	2,82	15,43	6,36
q_{Pe} = 10 biomarqueurs prédictifs +	FP / VP	1,22/0,82	2,49/5,69	1,73/1,09	9,53/5,91	4,40/1,96
q_{Po} = 10 biomarqueurs pronostiques	FDR / FNR	0,44/0,92	0,28/0,43	0,57/0,89	0,63/0,41	0,67/0,80

Résultats avec $p = 500$ biomarqueurs et $n = 1500$ patients. Quantités moyennes sur 500 réplications. q_{Pe} : nombre de biomarqueurs prédictifs, q_{Po} : nombre biomarqueurs pronostiques, n_{Po} : nombre des effets propres sélectionnés correspondant à des interactions sélectionnées, FP : faux positifs, VP : vrais positifs, FDR : taux de fausses découvertes, FNR : taux de faux négatifs, AL : lasso adaptatif, LRT : test du rapport de vraisemblance

FIGURE 4.2 : *Capacité de sélection des effets propres des biomarqueurs prédictifs (interactions) dans les scénarios alternatifs*



Résultats avec $p = 500$ *biomarqueurs et* $n = 1500$ *patients. Quantités moyennes sur 500 réplications.* q_{Pe} *: nombre de biomarqueurs prédictifs,* q_{Po} *: nombre biomarqueurs pronostiques,* FDR *: taux de fausses découvertes,* FNR *: taux de faux négatifs, AL : lasso adaptatif*

TABLEAU 4.4 : La concordance pour bénéfice et (l'erreur type empirique) des différentes méthodes dans les scénarios alternatifs

	AL(ridge)	AL(LRT)	cMCP	gel	SGL
Scénario 2 q_{Pe} = 1 biomarqueur prédictif	0,67 (0,02)	0,68 (0,02)	0,67 (0,02)	0,67 (0,02)	0,50 (0,03)
Scénario 3 q_{Pe} = 10 biomarqueurs prédictifs	0,88 (0,01)	0,88 (0,01)	0,88 (0,01)	0,87 (0,01)	0,53 (0,03)
Scénario 4 q_{Pe} = 10 biomarqueurs prédictifs + q_{Po} = 10 biomarqueurs pronostiques	0,72 (0,02)	0,72 (0,02)	0,72 (0,02)	0,72 (0,02)	0,52 (0,03)

Résultats avec $p = 500$ biomarqueurs et $n = 1500$ patients. Quantités moyennes sur 500 réplications. q_{Pe} : nombre de biomarqueurs prédictifs, q_{Po} : nombre biomarqueurs pronostiques, n_{Pe} : nombre d'interactions sélectionnées, FP : faux positifs, VP : vrais positifs, FDR : taux de fausses découvertes, FNR : taux de faux négatifs, AL : lasso adaptatif, LRT : test du rapport de vraisemblance

4.6 Application : cancer du sein précoce

Contexte

Afin d'illustrer les approches présentées dans ce travail de recherche, nous les avons appliquées à des données réelles issues de l'essai clinique B-31 du (*National Surgical Adjuvant Breast and Bowel Project (NSABP)*) évaluant l'effet du trastuzumab adjuvant chez des patientes atteintes d'un cancer du sein précoce. Un total de $n = 1574$ patientes étaient randomisées en deux bras de traitement : un bras chimiothérapie seule (bras C, $n = 795$) et un bras chimiothérapie et trastuzumab comme traitement adjuvant (bras C + T, $n = 779$) (L et al. [2013]). Le taux de censure est de 73% (soit 431 événements) et les taux de survie sans récidive à distance (SSRD) à 5 ans sont égales à 84% [95% IC : 81% -86%] et 64% [95% IC : 61% -68%] pour les patientes des bras C + T et C, respectivement. Le trastuzumab en association avec la chimiothérapie a amélioré significativement la survie sans récidive à distance par rapport à la chimiothérapie seule avec un $HR = 0{,}46$ [95% IC : 0,38 -0,56]. Néanmoins, cet effet peut ne pas être le même chez toute la population étudiée et les avantages de l'ajout de trastuzumab pourraient être déterminés par des interactions gènes-traitement. Les données d'expression génique avaient été recueillies pour $p = 462$ gènes. L'objectif de cette application est d'identifier des interactions gènes-traitement et les effets propres corresponds des gènes prédictifs de l'effet du trastuzumab chez les patientes diagnostiquées avec un cancer du sein.

Résultats

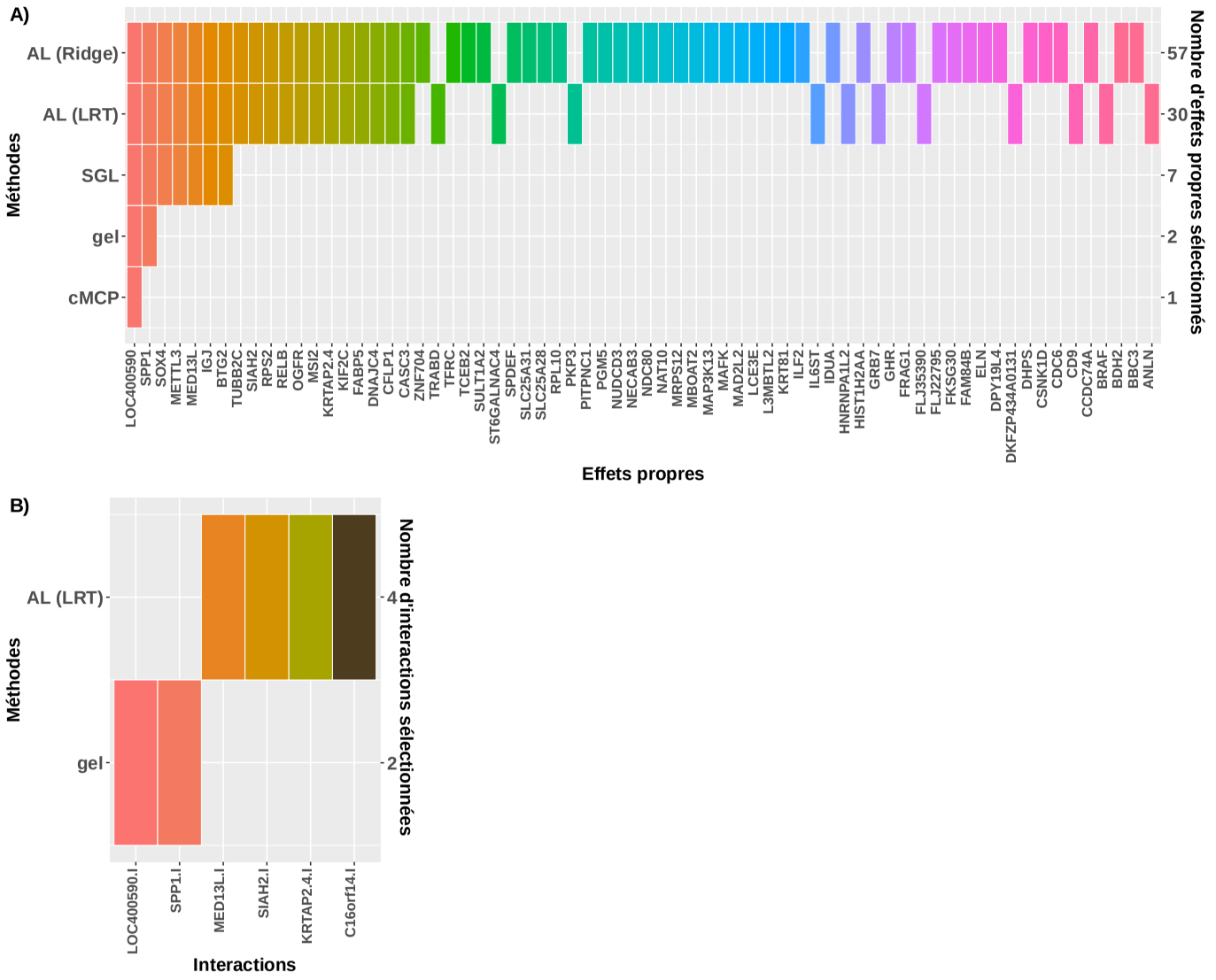
A cause de la variabilité des résultats engendrée par le processus de la validation croisée pour le choix du paramètre λ, nous avons effectué 100 répétitions. La figure 4.3 montre le résultat des effets propres (A) et des interac-

tions gènes-traitements (B) sélectionnés dans au moins 10% des 100 répétitions, ainsi que le nombre d'effets propres et des interactions sélectionnés par chaque méthode. Les résultats dans la figure (A) montrent qu'il existe une différence entre les méthodes quant au nombre d'effets propres sélectionnés. En effet, les méthodes lasso adaptatif avec les pondérations ridge et test de rapport de vraisemblance partielle (LRT) ont sélectionné un nombre plus élevé d'effets propres que les autres méthodes, à savoir 57 et 30, respectivement. La méthode cMCP n'a identifié qu'un seul gènes (LOC400590), la méthode gel a sélectionné deux effets propres (LOC400590, SPP1) et la méthode SGL a identifié 7 effets propres (LOC400590, SPP1, SOX4, METTL3, MED13L, IGJ, BTG2). Les résultats dans la figure (B) montrent que seules les méthodes lasso adaptatif (LRT) et gel ont identifié des interactions, alors que les autres méthodes n'ont pas réussi à détecter des interactions gènes-traitement. La méthode gel a sélectionné deux interactions (LOC400590, SPP1) avec leurs effets propres correspondant. Tandis que la méthode lasso adaptatif (LRT) a sélectionné 4 autres interactions dont 3 (MED13L, SIAH2, KRTAP2.4) avec leurs effets propres et une seule interaction (C16orf14) sans l'effet propre correspondant.

Des études récentes (Divya et F [2013], Yaqin et al. [2019], Liang et al. [2019]) ont conclu que le gène Phosphoprotéine 1 sécrétée (SPP1), également connu sous le nom d'ostéopontine (OPN), était un biomarqueur pronostique potentiel ou une nouvelle cible thérapeutique pour le cancer du sein. En particulier, une surexpression du gène SPP1 dans les tumeurs primaires s'est avérée associée à un risque de récidive chez les patientes atteintes d'un cancer du sein et présentant un statut récepteur à l'oestrogène positif ($ER+$) (Eremo et al. [2020]).
D'autre part, Jansen et al. [2009] ont montré que la protéine SIAH2 est un prédicteur de l'évolution de la maladie dans le cancer du sein ER-positif après un

traitement au tamoxifène. Notamment, l'ARN messager de SIAH2 est associée significativement aux taux de protéines ER des tumeurs primaires du sein. En revanche, CHAN et al. [2011] ont constaté que les niveaux de protéines SIAH2 étaient majoritairement régulés à la hausse dans le cancer du sein ER-négatif et qu'une surexpression du SIAH2 était associée à une survie défavorable.

FIGURE 4.3 : *Les effets propres et les interactions gènes-traitement identifiés par les méthodes avec les données de l'expression de gènes du cancer du sein pour la survie sans récidive à distance*



L'axe des x représente les gènes identifiés dans au moins 10% des 100 répétitions par chaque méthode. AL : lasso adaptatif, LRT : test du rapport de vraisemblance

4.7 Conclusion

L'effet du traitement dans les essais cliniques est souvent affecté par les caractéristiques des patients en particulier par des biomarqueurs. En réalité, certains patients bénéficient plus du traitement, alors que d'autres patients n'en bénéficient pas ou au contraire peuvent subir des effets délétères. Cette hétérogénéité de l'effet du traitement joue un rôle central dans le domaine de la médecine personnalisée et facilite le développement de thérapies ciblées ou adaptées. Les biomarqueurs de grande dimension tels que les données génomiques sont de plus en plus mesurés dans des essais cliniques randomisés. Par conséquent, il existe un intérêt croissant pour le développement des méthodes qui améliorent la capacité de détecter les interactions entre les biomarqueurs et le traitement. Par ailleurs, l'omission ou une mauvaise spécification des effets propres (variables de confusion) peuvent introduire un biais dans l'interaction (VanderWeele et al. [2012]). D'autres travaux (Cox [1984], Bien et al. [2013]) exigent qu'une interaction n'est incluse dans le modèle uniquement si les effets principaux pertinents sont également inclus dans le modèle.

L'objectif de ce travail de thèse est de favoriser la contrainte hiérarchique d'interaction biomarqueurs-traitement, c'est-à-dire favoriser la sélection des effets propres correspondant à des interactions sélectionnées par le modèle. Dans le cadre des essais cliniques randomisés et des données en grande dimension, nous avons proposé une approche de pondération pour la méthode lasso adaptatif basée sur le test de rapport de vraisemblance partielle (LRT) afin de favoriser la contrainte hiérarchique d'interaction. De fait, nous avons proposé de créer des groupes composés de l'effet propre du biomarqueur et de l'interaction biomarqueur-traitement, ensuite, d'attribuer à chaque groupe des poids adaptatifs basés sur le test de rapport de vraisemblance partielle (LRT).

Dans une étude de simulation couvrant plusieurs scénarios, nous avons évalué et comparé l'approche proposée lasso adaptatif (LRT) aux méthodes alternatives lasso adaptatif (ridge), cMCP, gel, et SGL en termes de performance sélective et prédictive. Globalement, les principaux résultats suggèrent que la méthode lasso adaptatif (LRT) favorise plus la contrainte hiérarchique d'interaction biomarqueurs-traitement en comparant avec les autres méthodes. En effet, l'approche proposée a sélectionné plus d'effets propres correspondant à des interactions sélectionnées que ces compétiteurs et elle avait la meilleure balance $FDR - FNR$ des effets propres sélectionnés. Les poids proposés $\Lambda_{M_2 - M_0}$ permettent d'accorder une faible pénalisation à l'effet propre d'un biomarqueur lorsque l'effet de son interaction avec le traitement est important. De cette façon, quelle que soit la force de cet effet propre, les chances qu'il soit sélectionné dans le modèle final sont plus élevées. Les méthodes cMCP et gel avaient de bons résultats pour l'identification des interactions mais de moins bons pour la sélection des effets propres correspondant. La méthode lasso adaptatif (ridge) était la plus conservatrice en terme de sélection des effets propres. Cependant, en terme de performance prédictive, les résultats étaient similaires entre les méthodes sauf la méthode SGL qui avait la plus faible valeur de la statistique de concordance pour bénéfice.

Dans cet axe de recherche, nous nous sommes restreints à favoriser la hiérarchie d'interaction biomarqueurs-traitement sans nécessairement forcer l'effet propre à être inclut dans le modèle final. Par conséquent, ce travail ne se situe pas dans le cadre d'une hiérarchie forte (*strong hierarchy*) comme celle présentée par Bien et al. [2013] entre biomarqueur-biomarqueur dans le cas du modèle linéaire. D'autre part, la statistique C-pour-bénéfice a été proposée récemment, elle fait encore objet de recherche (Thomas Debray [2020]) et nous connaissons peu les caractéristiques opérationnelles de cette mesure.

Ce travail effectué pourra être complété par une réévaluation de la force d'interaction entre les biomarqueurs et le traitement en calculant la différence des statistiques de concordance par bras de traitement ΔC (Ternes, Rotolo, Heinze et al. [2017]). Ce critère permettra d'évaluer la force prédictive en mesurant la différence de la valeur pronostique entre les deux bras de traitement. En outre, l'omission des effets propres peut donner une mauvaise calibration du modèle, ainsi, nous pourrons examiner l'effet de l'inclusion des effets propres sur la calibration du modèle qui est souvent visualisée au moyen de la courbe de calibration.

4.8 Références

Betensky, R. A., Louis, D. N. & Gregory Cairncross, J. (2002). Influence of unrecognized molecular heterogeneity on randomized clinical trials. *Journal of Clinical Oncology*, *20*(10), 2495-2499. https://doi.org/10.1200/JCO.2002.06.140. 113

Buyse, M. & Michiels, S. (2013). Omics-based clinical trial designs. *Current opinion in oncology*, *25*(3), 289-295. https://doi.org/10.1097/CCO.0b013e32835ff2fe. 113

Stendahl, M., Rydén, L., Nordenskjöld, B., Jönsson, P. E., Landberg, G. & Jirström, K. (2006). High progesterone receptor expression correlates to the effect of adjuvant tamoxifen in premenopausal breast cancer patients. *Clinical Cancer Research*, *12*(15), 4614-4618. https://doi.org/1078-0432.CCR-06-0248. 113

Delozier, T. (2010). Hormonothérapie du cancer du sein. *Journal de gynécologie obstétrique et biologie de la reproduction*, *39*(8), F71-F78. https://doi.org/10.1016/j.jgyn.2010.10.004. 113

(EBCTCG), E. B. C. T. C. G. (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen : patient-level meta-analysis of randomised trials. *The lancet*, *378*(9793), 771-784. https://doi.org/10.1016/S0140-6736(11)60993-8. 113

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17-36. https://doi.org/10.2307/3315687. 115

Hamada, M. & Wu, C. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of quality technology*, *24*(3), 130-137. https://doi.org/10.1080/00224065.1992.11979383. 115

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1-24. https://doi.org/10.2307/1403235. 115, 143

McCullagh, P. (2019). Generalized linear models. 115.

Bien, J., Taylor, J. & Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, *41*(3), 1111. https://doi.org/10.1214/13-AOS1096. 115, 143, 144

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, *34*(2), 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. 115

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429. https://doi.org/10.1198/016214506000000735. 116

Zhang, H. H. & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, *94*(3), 691-703. https://doi.org/10.1093/biomet/asm037. 116

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. 117

Ternes, N., Rotolo, F., Heinze, G. & Michiels, S. (2017). Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biometrical Journal*, *59*(4), 685-701. https://doi.org/10.1002/bimj.201500234. 121, 145

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67. 121, 122.

Breheny, P. & Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface*, *2*(3), 369-380. 123.

Huang, J., Breheny, P. & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science : a review journal of the Institute of Mathematical Statistics*, *27*(4). https://doi.org/10.1214/12-STS392. 123

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, *38*(2), 894-942. https://doi.org/10.1214/09-AOS729. 123

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, *71*(3), 731-740. https://doi.org/10.1111/biom.12300. 123, 129

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231-245. https://doi.org/10.1080/10618600.2012.681250. 123

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *68*(1), 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x. 124

van Klaveren, D., Steyerberg, E. W., Serruys, P. W. & Kent, D. M. (2018). The proposed concordance-statistic for benefitprovided a useful metric when modeling heterogeneous treatment effects. *Journal of clinical epidemiology*, *94*, 59-68. https://doi.org/10.1016/j.jclinepi.2017.10.021. 127

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011). Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of statistical software*, *39*(5), 1. https://doi.org/10.18637/jss.v039.i05. 129

Friedman, J., Hastie, T., Simon, N. & Tibshirani, R. (2018). *glmnet : Lasso and Elastic-Net Regularized Generalized Linear Models* [R-package version 2.0-16]. https://cran.r-project.org/web/packages/glmnet. 129

TERNES, N., ROTOLO, F. & MICHIELS, S. (2017). *biospear : Biomarker Selection in Penalized Regression Models* [R package version 1.0.1]. https://CRAN.R-project.org/package=biospear. 129

PATRICK, B. (2020). *grpreg : Regularization Paths for Regression Models with Grouped Covariates* [R package version 3.3.0]. https://CRAN.R-project.org/package=grpreg. 129

NOAH, S., JEROME, F., TREVOR, H. & ROB, T. (2019). *SGL : Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization* [R package version 1.3]. https://CRAN.R-project.org/package=SGL. 129

L, P.-G. K., CHUNGYEUL, K., JONG-HYEON, J., NORIKO, T., HANNA, B., G, G. P., DEBORA, F., C, G. L., NOUR, S., EIKE, B. et al. (2013). Predicting degree of benefit from adjuvant trastuzumab in NSABP trial B-31. *Journal of the National Cancer Institute, 105*(23), 1782-1788. https://doi.org/10.1093/jnci/djt321. 139

DIVYA, R. & F, W. G. (2013). An osteopontin promoter polymorphism is associated with aggressiveness in breast cancer. *Oncology reports, 30*(4), 1860-1868. https://doi.org/10.3892/or.2013.2632. 140

YAQIN, T., CAI, C. & GUORUN, F. (2019). Association between the expression of secreted phosphoprotein-related genes and prognosis of human cancer. *BMC cancer, 19*(1), 1-12. https://doi.org/10.1186/s12885-019-6441-3. 140

LIANG, L., LU, G., GUOGANG PAN, G., DENG, Y., LIANG, J., LIANG, L., LIU, J., TANG, Y. & WEI, G.-J. (2019). A case-control study of the association between the SPP1 gene SNPs and the susceptibility to breast cancer in Guangxi, China. *Frontiers in oncology, 9*, 1415. https://doi.org/10.3389/fonc.2019.01415. 140

Eremo, A. G., Lagergren, K., Othman, L., Montgomery, S., Andersson, G. & Tina, E. (2020). Evaluation of SPP1/osteopontin expression as predictor of recurrence in tamoxifen treated breast cancer. *Scientific reports, 10*(1), 1-9. https://doi.org/10.1038/s41598-020-58323-w. 140

Jansen, M. P., Ruigrok-Ritstier, K., Dorssers, L. C., van Staveren, I. L., Look, M. P., Meijer-van Gelder, M. E., Sieuwerts, A. M., Helleman, J., Sleijfer, S., Klijn, J. G. et al. (2009). Downregulation of SIAH2, an ubiquitin E3 ligase, is associated with resistance to endocrine therapy in breast cancer. *Breast cancer research and treatment, 116*(2), 263-271. https://doi.org/10.1007/s10549-008-0125-z. 140

Chan, P., Möller, A., Liu, M. C., Sceneay, J. E., Wong, C. S., Waddell, N., Huang, K. T., Dobrovic, A., Millar, E. K., O'Toole, S. A. et al. (2011). The expression of the ubiquitin ligase SIAH2 (seven in absentia homolog 2) is mediated through gene copy number in breast cancer and is associated with a basal-like phenotype and p53 expression. *Breast Cancer Research, 13*(1), 1-10. https://doi.org/10.1186/bcr2828. 141

VanderWeele, T. J., Mukherjee, B. & Chen, J. (2012). Sensitivity analysis for interactions under unmeasured confounding. *Statistics in Medicine, 31*(22), 2552-2564. https://doi.org/10.1002/sim.4354. 143

Thomas Debray, J. H., Jeroen Hoogland. (2020). An evaluation of the c-statistic for benefit [41st Annual Conference of the International Society for Clinical Biostatistics (ISCB), Krakow]. 144.

Chapitre 5

Conclusion générale et perspectives

Dans ce travail de recherche, nous nous sommes intéressés à l'identification des différentes classes de biomarqueurs dans les essais cliniques en oncologie. L'objectif de cette thèse a été de développer et d'évaluer de nouvelles méthodes statistiques adaptées aux exigences de validation des critères de substitutions de type survie d'une part et de sélection des biomarqueurs pronostiques et prédictifs dans le cadre des données en grande dimension d'autre part.

Dans la première partie de ce travail, nous avons proposé une méthode de validation d'un critère de substitution pour un critère de jugement principal de type survie au niveau de l'étude. Cette méthode est basée sur l'erreur de prédiction et représente une alternative au coefficient de détermination R^2 communément utilisé.
Les mesures de distance présentées sont estimées à l'aide de données provenant de plusieurs essais randomisés comme une méta-analyse. Elles permettent de mesurer la force de la corrélation entre le critère de substitution et le critère de jugement principal de type survie en termes d'erreur de prédiction, autre-

ment dit, en fonction de la différence moyenne entre l'effet du traitement prédit et observé sur le critère de jugement principal.

Il est utile de noter que la validation d'un biomarqueur de substitution nécessite la présence d'une forte association entre les deux critères de jugement à la fois au niveau individuel et au niveau de l'étude, en d'autres termes, une forte association entre les effets du traitement sur le biomarqueur et sur le critère de jugement principal.

Cependant, la force de l'association au niveau individuel est une condition nécessaire mais elle n'est pas suffisante pour qu'un biomarqueur soit considéré comme un substitut valide. L'association au niveau individuel est souvent simple à mettre en évidence et ne nécessite en principe que des données provenant d'un seul essai. L'association au niveau de l'étude nécessite plusieurs essais et est souvent plus difficile à mettre en évidence de façon convaincante.

Dans une étude de simulation nous avons étudié deux scénarios principaux (un petit nombre de grands essais et un grand nombre de petits essais) que nous avons considéré informatifs et ne nécessitent pas de calculs intensifs. Bien que les résultats de cette étude soient cohérents dans tous les scénarios, il est difficile de recommander un seuil pour la décision finale, tout comme pour le coefficient R^2.

Il est important de souligner que les distances moyennes proposées permettent d'illustrer l'erreur de prédiction pour un effet du traitement estimé en matière de valeurs moyennes pour la sous-prédiction ou sur-prédiction des effets du traitement. Mesurées sur l'échelle de l'effet du traitement, elles fournissent des informations supplémentaires importantes sur la qualité de prédiction et constituent une mesure plus stable et facile à interpréter par le clinicien que le coefficient R^2.

En ce qui concerne la sélection des biomarqueurs pronostiques et prédictifs, nous nous sommes focalisés essentiellement sur des méthodes de régressions pénalisées en particulier la méthode lasso adaptatif qui est une variante de la méthode lasso standard. Le choix de cette méthode se justifie notamment par les atouts qu'elle présente, à savoir : le lasso adaptatif possède la propriété d'oracle et de plus cette méthode est flexible puisqu'elle permet d'attribuer des poids différents à chaque biomarqueur, ce qui constitue un avantage utile dans le processus de sélection.

Dans le cadre pronostique, notre objectif était de prendre en compte l'information disponible sur l'appartenance des biomarqueurs à des groupes prés-pécifiés comme les voies biologiques (*pathways*). À cet effet, nous souhaitions privilégier l'identification des gènes potentiellement pronostiques appartenant à des voies qui montrent un rôle pronostique. Pour cela, nous avons proposé huit stratégies de pénalisation pour la méthode lasso adaptatif. Ces poids peuvent être regroupés en deux familles de pondération. La première est basée sur le coefficient de régression univariée de Cox et la deuxième famille de poids est basée sur la statistique de Wald univariée. Ces différentes pondérations proposées cherchent à effectuer la sélection à deux niveaux : à la fois des groupes les plus pronostiques et des biomarqueurs pertinents au sein des groupes sélectionnés, tout en éliminant les faux positifs. Pour montrer les caractéristiques opérationnelles de nos approches basées sur le lasso adaptatif, nous les avons comparées à d'autres méthodes alternatives permettant la sélection à deux niveaux dans une étude de simulation à huit scénarios. Les résultats de cette étude ont montré que les approches de pondération proposées pour le lasso adaptatif avec double poids au niveau individuel et du groupe étaient les plus parcimonieuses en termes de nombre de biomarqueurs sélectionnés. De plus, elles avaient les meilleurs compromis des taux de fausses découvertes et de

faux négatifs. Par ailleurs, dans ce travail, nous avons considéré des groupes de biomarqueurs non corrélés et disjoints ; il serait intéressant d'étendre les approches proposées au cas des groupes qui se chevauchent par duplication des biomarqueurs.
Bien qu'il existe de nombreuses approches possibles incorporant la connaissance a priori sur la structure en groupe des biomarqueurs comme l'approche bayésienne (Tang et al. [2019]), les poids basés sur la statistique de Wald proposés expriment une vision différente de pondération pour la méthode lasso adaptatif. Notamment, l'approche proposée prend en compte l'appartenance de chaque biomarqueur à un groupe, mais les poids des différents groupes sont basés sur le rôle pronostique observé dans les données et non pas sur une connaissance ou une croyance à priori.

Concernant les biomarqueurs prédictifs de l'effet du traitement ou les modificateurs de l'effet du traitement, nous nous sommes intéressés à favoriser la contrainte hiérarchique entre les interactions biomarqueur-traitement et leurs effets propres dans la procédure de sélection des biomarqueurs prédictifs dans le cadre de la survie. Nous avons proposé une reparamétrisation du modèle de Cox d'interaction hiérarchique biomarqueurs-traitement en créant des groupes composés de l'effet propre du biomarqueur et son interaction avec le traitement. Par la suite, nous avons développé des pondérations pour la méthode lasso adaptatif basées sur le test de rapport de vraisemblance partielle dans le but de pénaliser chaque groupe différemment tout en encourageant la sélection des interactions biomarqueurs-traitement pertinentes avec leurs effets propres associés. L'étude de simulation a montré que globalement la méthode lasso adaptatif avec la pondération par la statistique du test de rapport de vraisemblance partielle permet de favoriser la contrainte hiérarchique des interac-

tions biomarqueurs-traitement en comparaison avec d'autres méthodes alternatives, mais au prix d'une plus grande probabilité de sélectionner davantage de faux biomarqueurs. Pour évaluer la capacité prédictive des méthodes, nous avons utilisé la mesure de discrimination concordance pour bénéfice. Bien que cette mesure nous a permis d'évaluer la capacité du modèle à distinguer les patients susceptibles de bénéficier davantage du traitement de ceux qui en bénéficient moins, nous connaissons peu ses caractéristiques opérationnelles. Ainsi, il sera important de réévaluer la force d'interaction entre les biomarqueurs et le traitement en calculant la différence des statistiques de concordance par bras de traitement. En outre, la prise en compte ou l'omission des effets propres peut avoir des conséquences sur le biais d'estimations et la calibration du modèle. Comme perspective, nous pourrions examiner la courbe de calibration pour bénéfice qui permet d'évaluer si le bénéfice du traitement absolu observé correspond au bénéfice prédit.
D'autre part, à notre connaissance, il n'existe pas de modèle statistique de sélection de biomarqueurs prédictifs permettant de respecter strictement la contrainte hiérarchique d'interaction biomarqueurs-traitement dans le cadre de la survie. L'approche que nous avons proposée peut être considérée comme une représentation des avantages et des inconvénients de la pénalisation des effets propres et des interactions.

Une autre perspective à ce travail de recherche est l'intégration à la fois des données provenant de plusieurs sources (cliniques, génomiques, etc.) dans les modèles de sélection et de prédiction. Souvent, il n'est pas nécessaire de faire la sélection sur les variables connues comme les facteurs socio-démographiques (comme le genre et l'âge, etc.) et cliniques (comme le statut des récepteurs hormonaux, le grade tumoral, et la taille de la tumeur, etc.). Toutefois, il est in-

téressant de noter que la méthode lasso adaptatif nous permet d'intégrer ces variables, que nous ne souhaitons pas pénaliser, en leur attribuant une pondération nulle, ce qui garantit leur sélection dans le modèle final. En revanche, il faut étudier l'impact de l'inclusion de ces variables sur la qualité de prédiction.

Dans ce travail de thèse, nous avons cherché à favoriser la contrainte hiérarchique d'interaction sans réellement forcer l'effet propre associé à être inclus dans le modèle final. Cependant, la question de l'inclusion des effets propres en présence d'interaction entre deux variables fait toujours l'objet de débats dans la littérature statistique. Les recherches ont amené à deux pensées : la première est la hiérarchie faible (ou *weak hierarchy*) qui stipule que si l'interaction est incluse dans le modèle alors l'un des deux effets propres relatifs doit l'être aussi. En revanche, la seconde est la hiérarchie forte (ou *strong hierarchy*) qui énonce que si une interaction est sélectionnée alors les deux effets propres doivent être obligatoirement inclus dans le modèle (Bien et al. [2013]). Une idée que nous trouvons intéressante est d'utiliser plusieurs stratégies de pondérations destinées au lasso adaptatif et basées sur la statistique de Wald, dans la même optique que le deuxième axe de thèse, en vue de respecter la contrainte hiérarchique faible et forte. Seulement, la question qui se pose : qu'en est-il des conséquences du choix d'intégrer ou pas les effets propres dans le modèle sur la qualité de prédiction en termes de discrimination et de calibration pour le bénéfice ?

En conclusion, dans ce travail de thèse, nous avons proposé de nouvelles approches statistiques pour valider un critère de substitution et pour la sélection des biomarqueurs pronostiques et prédictifs dans un contexte de grande dimension pour l'analyse des données censurées. Le code `R` implémentant les

approches proposées est disponible sur demande et sera intégré dans le package `biospear` prochainement.

5.1 Références

TANG, Z., LEI, S., ZHANG, X., YI, Z., GUO, B., CHEN, J. Y., SHEN, Y. & YI, N. (2019). Gsslasso Cox : a Bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information. *BMC Bioinformatics, 20*(1), 94. https://doi.org/10.1186/s12859-019-2656-1. 154

BIEN, J., TAYLOR, J. & TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *Annals of statistics, 41*(3), 1111. https://doi.org/10.1214/13-AOS1096. 156

Annexes

Annexe A

Résultats complémentaires pour la validation d'un critère de substitution

TABLEAU A.1 : Exemple de critères de substitution acceptés par la FDA (*Food and Drug Administration*)

Maladie ou utilisation	**Population de patients**	**Critères de substitution**	**Type d'homologation approprié**	**Mécanisme d'action du médicament**
Cancer : tumeurs solides	Les patientes atteintes d'un cancer du sein ; cancer des ovaires ; carcinome rénal ; le cancer neuroendocrinien du pancréas ; cancer colorectal ; cancer de la tête et du cou ; cancer du poumon non à petites cellules ; cancer du poumon à petites cellules ; mélanome ; complexe de sclérose tubéreuse ; Carcinome à cellules de Merkel ; carcinome basocellulaire ; carcinome urothélial ; cancer du col de l'utérus ; cancer de l'endomètre ; carcinome hépatocellulaire ; cancer de la trompe de Fallope ; cancer à forte instabilité microsatellitaire ; cancer de l'estomac ; cancer de la thyroïde ; astrocytome ; Le sarcome de Kaposi lié au SIDA ; carcinome urotherlial ; carcinome épidermoïde cutané non résécable ou métastatique ; fusion du gène de la tyrosine kinase (NTRK) du récepteur neurotrophique sans mutation de résistance acquise connue	Taux de réponse objectif durables (*Durable objective response rates* ou ORR)	Accéléré / traditionnel	Indépendant du mécanisme
Cancer : tumeurs solides	Les patientes atteintes d'un cancer du sein ; carcinome rénal ; tumeur neuroendocrine pancréatique ; sarcome des tissus mous ; cancer de l'ovaire, de la trompe de Fallope ou du péritoine primaire ; cancer de la prostate ; cancer de la thyroïde ; cancer colorectal ; cancer du poumon non à petites cellules ; cancer de la tête et du cou ; complexe de sclérose tubéreuse ; Carcinome à cellules de Merkel ; carcinome basocellulaire ; carcinome urothélial ; cancer du col de l'utérus ; cancer de l'endomètre ; carcinome hépatocellulaire ; cancer de la trompe de Fallope ; mélanome ; astrocytome ; tumeurs stromales gastro-intestinales	Survie sans progression (*Progression-free survival* ou PFS)	Accéléré / traditionnel	Indépendant du mécanisme
Cancer : tumeurs solides	Patients recevant un traitement adjuvant après une résection chirurgicale complète du cancer du côlon ; cancer colorectal ; mélanome ; cancer des cellules rénales ; tumeur stromale gastro-intestinale ; cancer du sein et traitement adjuvant du cancer du poumon non à petites cellules de stade III	Survie sans maladie (*Disease-free survival* ou DFS)	Accéléré / traditionnel	Indépendant du mécanisme
Cancer : tumeurs solides	Les patientes atteintes d'un cancer du sein ; neuroblastome	Survie sans événement (*Event-free survival* ou EFS)	Accéléré / traditionnel	Indépendant du mécanisme
Cancer : tumeurs solides	Patients atteints d'un cancer du sein	Réponse complète pathologique	Accéléré	Indépendant du mécanisme
Cancer : tumeurs solides	Patients atteints d'un cancer de la prostate non métastatique résistant à la castration	Survie sans métastases	Accéléré / traditionnel	Indépendant du mécanisme
Cancer : tumeurs solides	Patients atteints d'un cancer de la prostate avancé	Niveaux de testostérone plasmatique	Traditionnel	Antagoniste de l'hormone de libération de la gonadotrophine

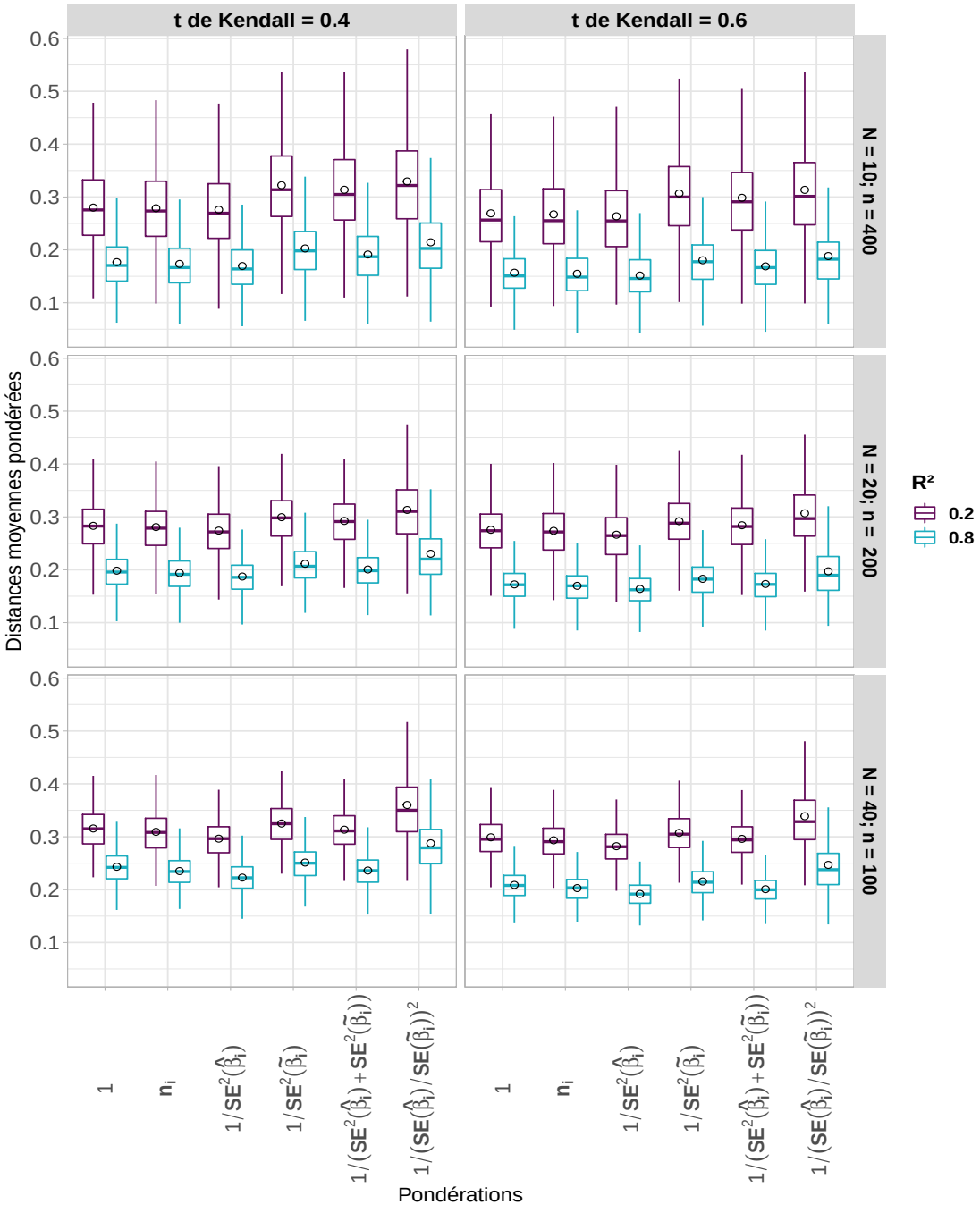
Tableau recréé à partir du tableau de la FDA sur https://www.fda.gov/drugs/development-resources/table-surrogate-endpoints-were-basis-drug-approval-or-licensure

TABLEAU A.2 : *Distances moyennes pondérées ($\overline{diff}(w_i)$), erreurs types empiriques (ETE) et erreurs types moyennes (ETM) avec les données générées avec la copule de Clayton non ajustée*

		scenario 1 ($\tau = 0.4; R^2 = 0.2$)			**scenario 2** ($\tau = 0.6; R^2 = 0.2$)			**scenario 3** ($\tau = 0.4; R^2 = 0.8$)			**scenario 4** ($\tau = 0.6; R^2 = 0.8$)		
	w_i	$\overline{diff}(w_i)$	ESE	(ASE)	$\overline{diff}(w_i)$	ESE	(ASE)	$\overline{diff}(w_i)$	ESE	(ASE)	$\overline{diff}(w_i)$	ESE	(ASE)
	1	0.280	0.072	(0.204)	0.269	0.076	(0.196)	0.177	0.051	(0.131)	0.157	0.044	(0.116)
	n_i	0.279	0.074	(0.192)	0.267	0.078	(0.184)	0.173	0.050	(0.122)	0.155	0.044	(0.108)
$N = 10$;	$1/SE^2(\hat{\beta}_{(i)})$	0.276	0.075	(0.189)	0.264	0.077	(0.181)	0.169	0.049	(0.119)	0.151	0.043	(0.106)
$n = 400$	$1/SE^2(\tilde{\beta}_{(i)})$	0.322	0.085	(0.211)	0.307	0.088	(0.200)	0.203	0.058	(0.134)	0.180	0.052	(0.119)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.314	0.082	(0.208)	0.299	0.086	(0.198)	0.191	0.056	(0.130)	0.169	0.048	(0.115)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.329	0.103	(0.209)	0.314	0.099	(0.198)	0.214	0.074	(0.135)	0.188	0.069	(0.119)
	1	0.283	0.050	(0.210)	0.276	0.051	(0.206)	0.198	0.037	(0.152)	0.172	0.033	(0.131)
	n_i	0.281	0.051	(0.202)	0.274	0.051	(0.199)	0.194	0.036	(0.144)	0.170	0.033	(0.126)
$N = 20$;	$1/SE^2(\hat{\beta}_{(i)})$	0.274	0.050	(0.197)	0.266	0.051	(0.193)	0.187	0.035	(0.139)	0.164	0.031	(0.121)
$n = 200$	$1/SE^2(\tilde{\beta}_{(i)})$	0.299	0.052	(0.215)	0.292	0.052	(0.211)	0.211	0.039	(0.157)	0.183	0.035	(0.135)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.292	0.051	(0.210)	0.284	0.052	(0.206)	0.201	0.037	(0.149)	0.173	0.033	(0.128)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.313	0.067	(0.219)	0.307	0.071	(0.217)	0.230	0.061	(0.166)	0.197	0.052	(0.142)
	1	0.316	0.040	(0.245)	0.299	0.038	(0.232)	0.243	0.032	(0.194)	0.209	0.027	(0.166)
	n_i	0.309	0.040	(0.237)	0.293	0.038	(0.223)	0.235	0.031	(0.184)	0.203	0.026	(0.158)
$N = 40$;	$1/SE^2(\hat{\beta}_{(i)})$	0.296	0.037	(0.225)	0.282	0.036	(0.213)	0.223	0.029	(0.172)	0.192	0.025	(0.147)
$n = 100$	$1/SE^2(\tilde{\beta}_{(i)})$	0.325	0.041	(0.250)	0.307	0.039	(0.236)	0.251	0.033	(0.199)	0.215	0.029	(0.171)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.313	0.039	(0.239)	0.296	0.037	(0.225)	0.236	0.031	(0.184)	0.201	0.026	(0.156)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.360	0.074	(0.274)	0.339	0.075	(0.258)	0.288	0.063	(0.225)	0.247	0.061	(0.195)

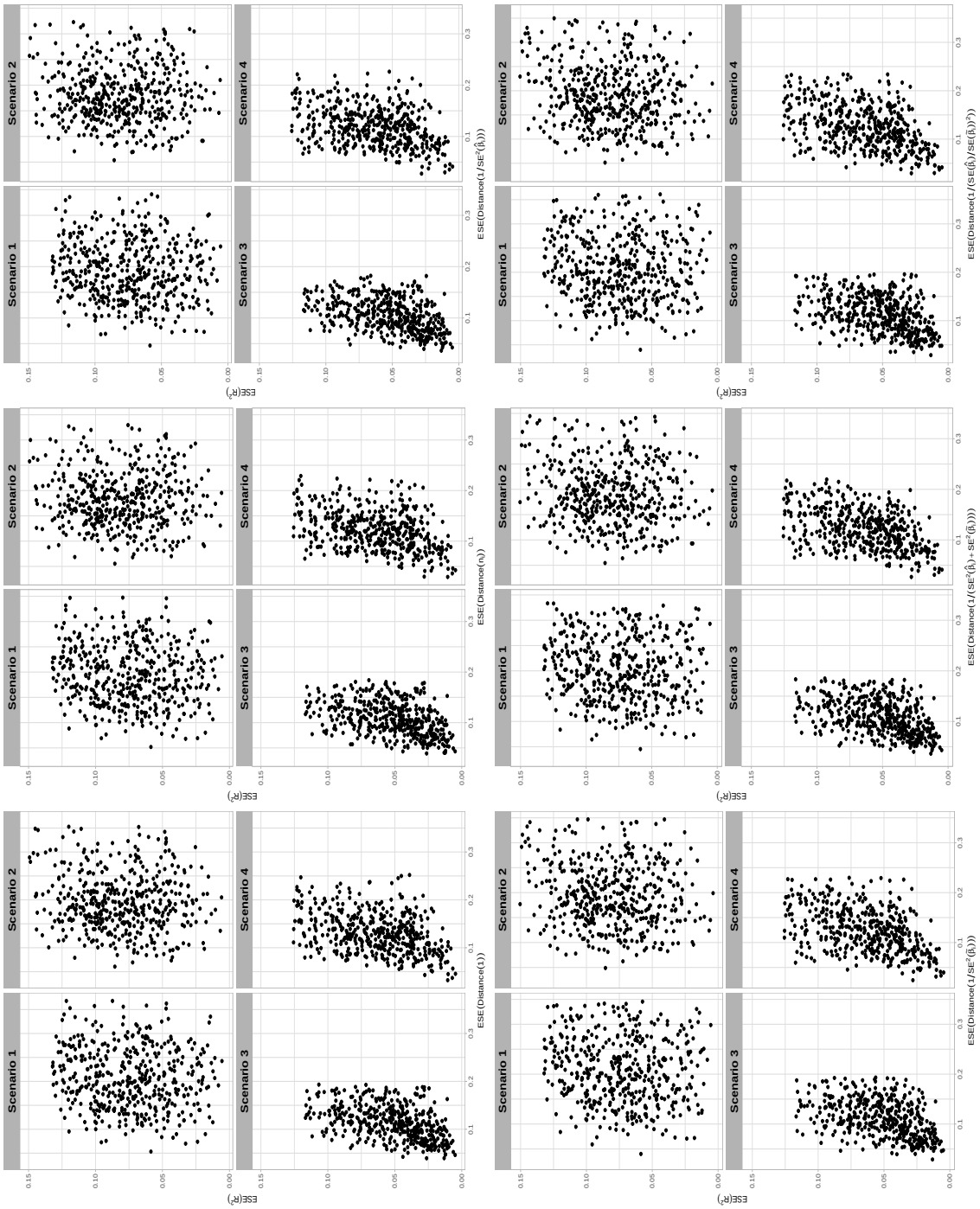
Légende. N : nombre d'essais, n : taille moyenne de l'essai. Quantités moyennes basées sur 500 réplications.

FIGURE A.1 : *Boxplots de distances moyennes pondérées selon les différents scénarios avec les données générées avec la copule de Clayton non ajustée*



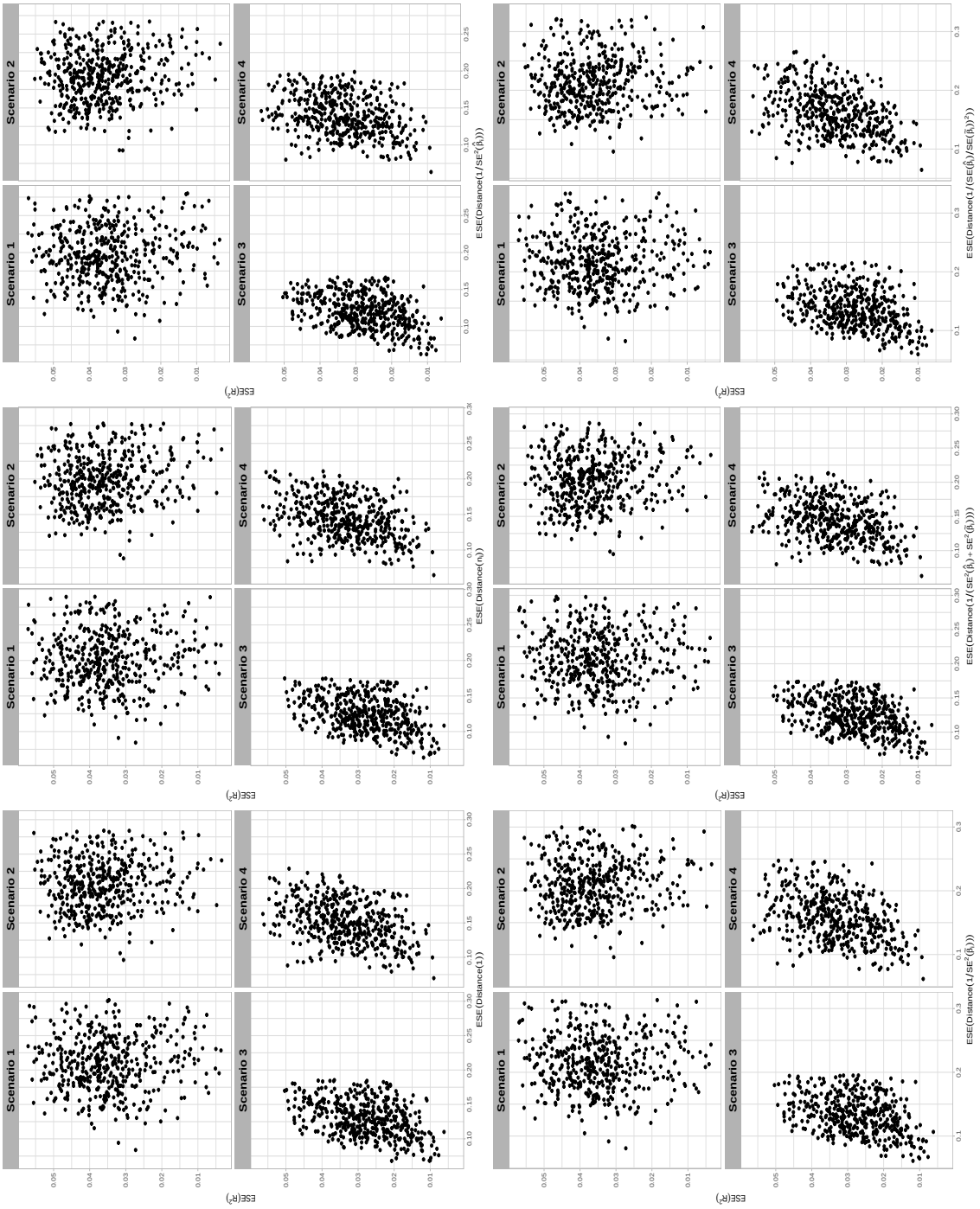
Légende. La boîte est délimitée par les 25e et 75e percentiles, contient 50% des observations et divisée par la médiane (trait horizontal à l'intérieur de la boîte). Les moustaches de style Tukey (traits fins horizontaux) s'étendent jusqu'à un maximum de 1, 5 *de l'intervalle interquartile au-delà de la boîte. Les cercles noirs sont les moyennes des distances sur 500 réplications. N : nombre d'essais, n : taille moyenne des essais*

FIGURE A.2 : Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées



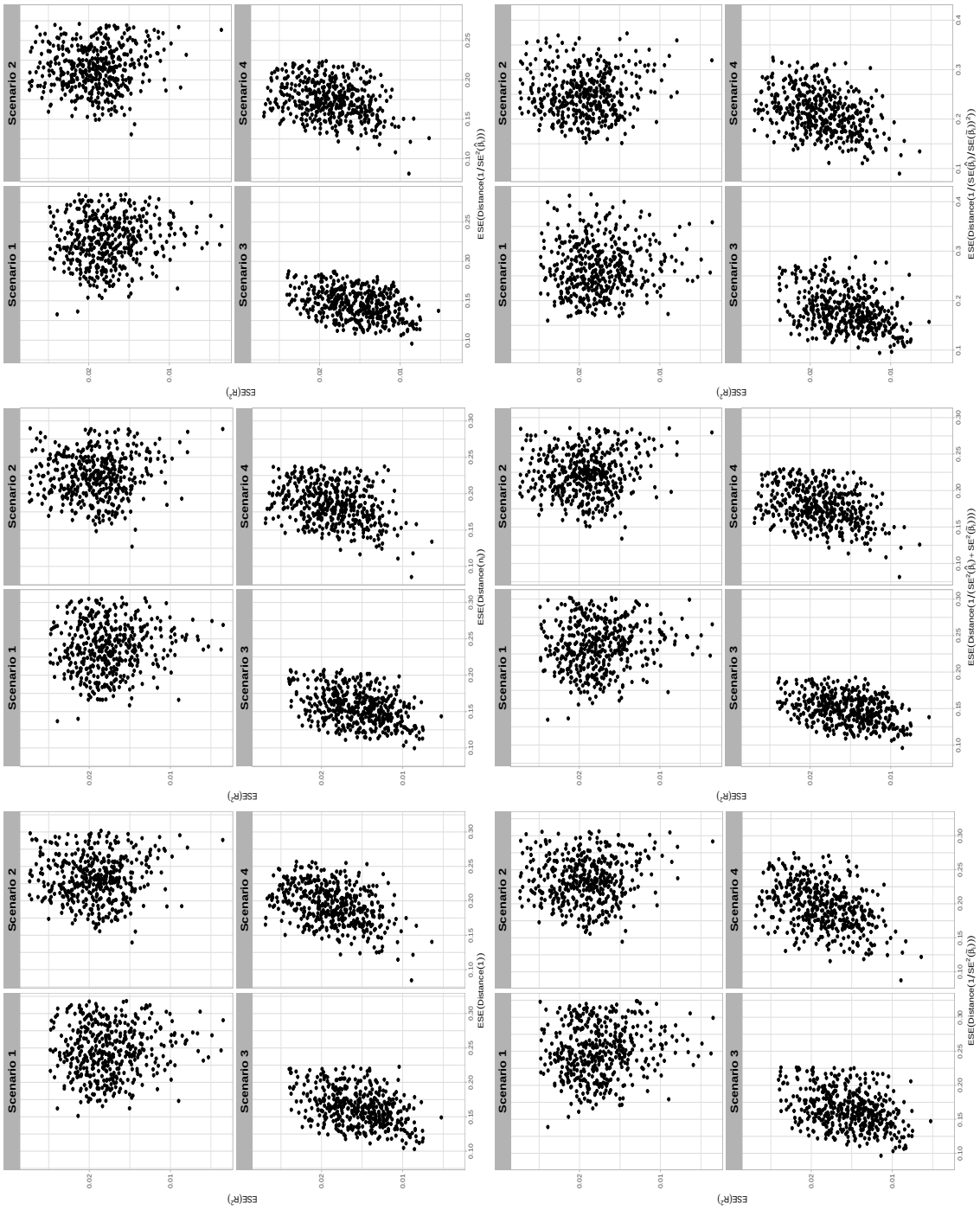
Résultats avec le modèle de copule de Clayton ajustée, N = 10 essais et de taille moyenne n = 400. ESE : erreur type empirique, Distance : distance moyenne pondérée, SE : erreur type

FIGURE A.3 : Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées



Résultats avec le modèle de copule de Clayton ajustée, N = 20 essais et de taille moyenne n = 200. ESE : erreur type empirique, Distance : distance moyenne pondérée, SE : erreur type

FIGURE A.4 : Nuage de points de l'erreur type empirique (ETE) du coefficient de détermination R^2 et des distances moyenne pondérées



Résultats avec le modèle de copule de Clayton ajustée, N = 40 essais et de taille moyenne n = 100. ESE : erreur type empirique, Distance : distance moyenne pondérée, SE : erreur type

Annexe B

Résultats complémentaires pour la sélection des biomarqueurs pronostiques dans le contexte de grande dimension

TABLEAU B.1 : *Conception de l'étude de simulation*

500 répétitions par scénario		**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
		Scénario 1 ($l = 0; q = 0$)	**Scénario 2** ($l = 20; q = 100$)	**Scénario 3** ($l = 1; q = 8$)	**Scénario 4** ($l = 2; q = 8$)	**Scénario 5** ($l = 1; q = 32$)	**Scénario 6** ($l = 2; q = 32$)	**Scénario 7** ($l = 2; q = 16$)	**Scénario 8** ($l = 2; q = 8$)
			HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
Censoring rate		78%	55%	66%	66%	71%	71%	74%	76%
(a) $p = 1000$ biomarqueurs;	AG Id	0	(1,...,20)	1	(1,2)	1	(1,2)	(1,2)	(1,2)
$R = 20$ groupes de taille égale 50	AB Id	0	(1,...,5, 51,...,55, 101,...,105,151,...,155, 201,...,205,251,...,255, 301,...,305,351,...,355, 401,...,405,451,...,455, 501,...,505,551,...,555, 601,...,605,651,...,655, 701,...,705,751,...,755, 801,...,805,851,...,855, 901,...,905,951,...,955)	(1,...,8)	(1,...,4, 51,...,54)	(1,...,32)	(1,...,16, 51,...,66)	(1,...,8, 51,...,58)	(1,...,4, 51,...,54)
(b) $p = 1000$ biomarqueurs;	AG Id	0	(1,...,20)	11	(1,11)	11	(1,11)	(1,11)	(1,11)
$R = 20$ groupes : 10 de taille 25 et 10 de taille 75	AB Id	0	(1,...,5, 26,...,30, 51,...,55, 76,...,80, 101,...,105,126,...,130, 151,...,155,176,...,180, 201,...,205,226,...,230, 251,...,255,326,...,330, 401,...,405,476,...,480, 551,...,555,626,...,630, 701,...,705,776,...,780, 851,...,855,926,...,930)	(251,...,258)	(1,...,4, 251,...,254)	(251,...,282)	(1,..,16, 251,...,266)	(1,...,8, 251,...,258)	(1,...,4, 251,...,254)

p : nombre total de biomarqueurs, R : nombre total de groupes de biomarqueurs, l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AG Id : index des groupes actifs, AB Id : index des biomarqueurs actifs

TABLEAU B.2 : *Résultats de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)*

Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 ($l = 0; q = 0$)	**Scénario 2 ($l = 20; q = 100$)**	**Scénario 3 ($l = 1; q = 8$)**	**Scénario 4 ($l = 2; q = 8$)**	**Scénario 5 ($l = 1; q = 32$)**	**Scénario 6 ($l = 2; q = 32$)**	**Scénario 7 ($l = 2; q = 16$)**	**Scénario 8 ($l = 2; q = 8$)**
		HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
Lasso standard	1,00/- (11,62)	0,48/0,13 (167,82)	0,81/0,04 (43,93)	0,81/0,04 (43,48)	0,71/0,49 (58,78)	0,71/0,50 (58,32)	0,76/0,54 (37,10)	0,80/0,57 (23,83)
Lasso adaptatif								
AC	1,00/- (15,10)	0,49/0,15 (167,54)	0,59/0,02 (21,96)	0,70/0,04 (28,94)	0,38/0,38 (34,54)	0,55/0,42 (43,71)	0,73/0,47 (35,42)	0,82/0,53 (28,42)
PCA	1,00/- (9,12)	0,59/0,53 (116,40)	0,64/0,05 (28,05)	0,73/0,11 (34,30)	0,50/0,46 (35,37)	0,59/0,54 (38,48)	0,72/0,62 (25,05)	0,81/0,70 (17,66)
Lasso + PCA	1,00/- (8,27)	0,50/0,18 (164,98)	0,56/0,03 (19,92)	0,64/0,09 (25,21)	0,60/0,49 (42,57)	0,62/0,54 (42,14)	0,70/0,57 (26,50)	0,76/0,61 (17,43)
SW	1,00/- (53,46)	0,13/0,15 (97,91)	0,29/0,02 (12,72)	0,36/0,03 (15,84)	0,58/0,45 (45,46)	0,57/0,45 (44,39)	0,80/0,44 (48,99)	0,90/0,44 (50,71)
ASW	1,00/- (17,10)	0,50/0,18 (165,17)	0,51/0,02 (16,72)	0,60/0,03 (20,58)	0,23/0,36 (27,24)	0,42/0,40 (34,31)	0,63/0,45 (27,67)	0,77/0,52 (24,44)
ASW × SW	1,00/- (49,62)	0,17/0,19 (98,36)	0,06/0,02 (8,35)	0,11/0,03 (8,87)	0,13/0,39 (24,16)	0,28/0,39 (29,90)	0,68/0,36 (37,98)	0,87/0,38 (43,04)
MSW	1,00/- (16,77)	0,52/0,24 (158,03)	0,50/0,02 (16,52)	0,61/0,03 (20,88)	0,32/0,37 (31,73)	0,45/0,40 (36,48)	0,62/0,45 (26,42)	0,75/0,53 (20,85)
MSW × SW	1,00/- (45,75)	0,18/0,21 (95,77)	0,06/0,02 (8,35)	0,11/0,03 (8,80)	0,24/0,38 (28,20)	0,30/0,39 (30,43)	0,64/0,35 (34,62)	0,83/0,36 (38,89)
cMCP	0,32/- (0,98)	0,10/0,69 (35,01)	0,41/0,58 (7,64)	0,42/0,60 (7,62)	0,31/0,75 (12,64)	0,35/0,77 (12,32)	0,38/0,79 (6,59)	0,34/0,81 (3,50)
gel	0,25/- (7,11)	0,05/0,52 (50,68)	0,19/0,04 (9,99)	0,16/0,05 (9,18)	0,31/0,13 (43,72)	0,19/0,80 (17,99)	0,11/0,86 (8,68)	0,10/0,85 (4,92)
SGL	0,40/- (8,99)	0,49/0,05 (185,16)	0,74/0,01 (42,36)	0,77/0,01 (45,88)	0,66/0,39 (65,84)	0,66/0,36 (66,67)	0,66/0,35 (41,26)	0,65/0,34 (28,36)
IPF-Lasso								
IPF-Lasso1	0,64/- (2,89)	0,38/0,16 (136,20)	0,42/0,03 (14,63)	0,45/0,04 (15,39)	0,23/0,44 (24,07)	0,30/0,46 (25,74)	0,41/0,50 (15,01)	0,50/0,56 (8,68)
IPF-Lasso2	0,64/- (2,89)	0,38/0,16 (136,20)	0,53/0,04 (18,50)	0,55/0,05 (19,12)	0,51/0,54 (31,57)	0,49/0,55 (29,73)	0,50/0,61 (14,86)	0,50/0,67 (7,92)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9)))

IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9)))

TABLEAU B.3 : *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Résultats de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de groupes de biomarqueurs sélectionnés)*

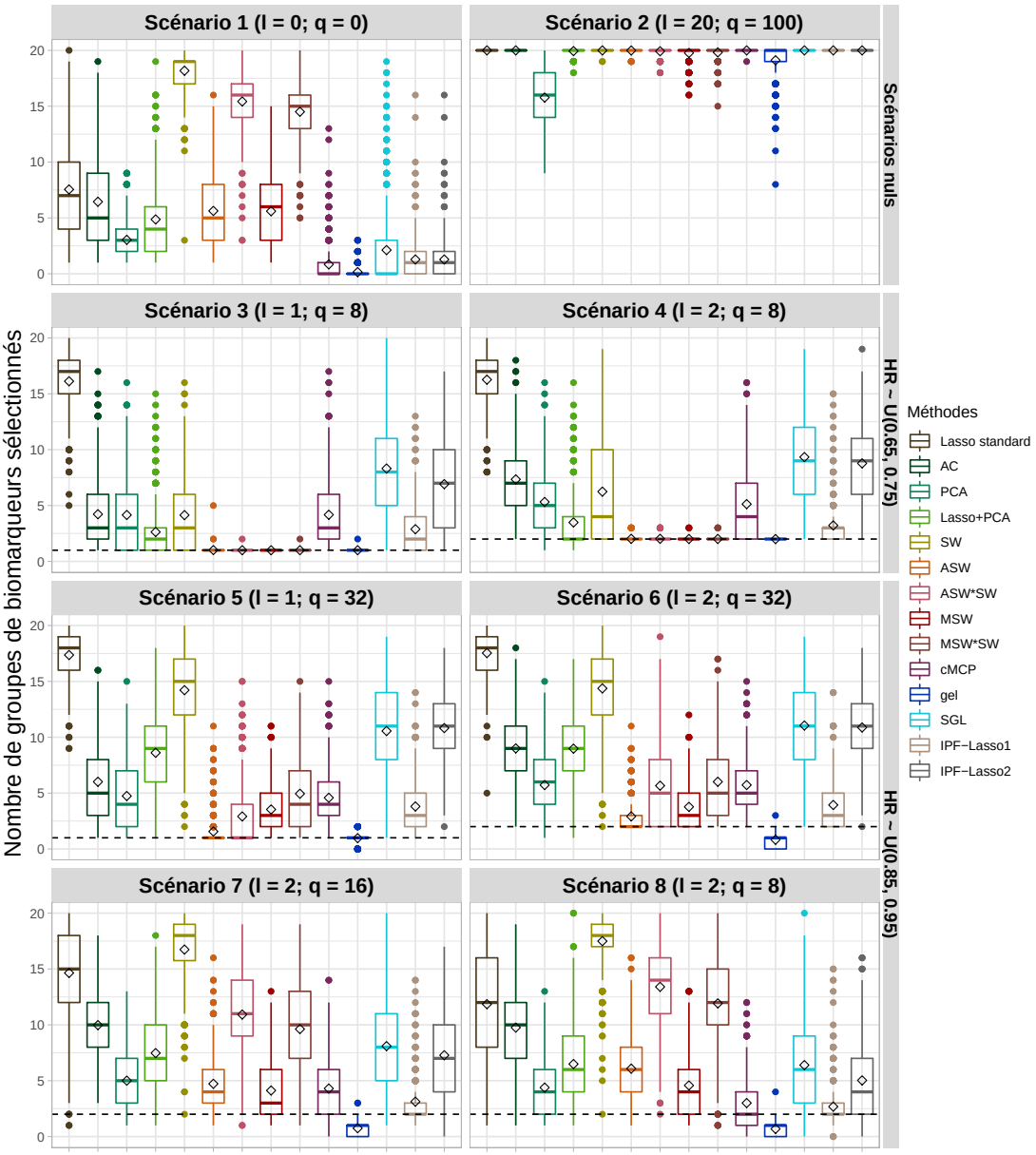
Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 ($l = 0; q = 0$)	**Scénario 2** ($l = 20; q = 100$)	**Scénario 3** ($l = 1; q = 8$)	**Scénario 4** ($l = 2; q = 8$)	**Scénario 5** ($l = 1; q = 32$)	**Scénario 6** ($l = 2; q = 32$)	**Scénario 7** ($l = 2; q = 16$)	**Scénario 8** ($l = 2; q = 8$)
		HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
Lasso standard	1,00/- (7,55)	-/0,00 (20,00)	0,94/0,00 (16,13)	0,87/0,00 (16,27)	0,94/0,00 (17,36)	0,88/0,00 (17,54)	0,84/0,00 (14,65)	0,80/0,04 (11,85)
Lasso adaptatif								
AC	1,00/- (6,45)	-/0,00 (20,00)	0,58/0,00 (4,22)	0,67/0,00 (7,34)	0,75/0,00 (6,02)	0,75/0,00 (9,00)	0,77/0,00 (9,98)	0,76/0,05 (9,76)
PCA	1,00/- (3,04)	-/0,21 (15,78)	0,53/0,02 (4,16)	0,52/0,06 (5,32)	0,71/0,10 (4,74)	0,63/0,16 (5,72)	0,63/0,26 (5,01)	0,65/0,39 (4,39)
Lasso + PCA	1,00/- (4,87)	-/0,00 (19,94)	0,35/0,01 (2,63)	0,29/0,05 (3,49)	0,88/0,07 (8,61)	0,78/0,11 (8,98)	0,72/0,12 (7,48)	0,67/0,18 (6,49)
SW	1,00/- (18,18)	-/0,00 (20,00)	0,49/0,00 (4,14)	0,43/0,00 (6,24)	0,92/0,00 (14,22)	0,84/0,00 (14,37)	0,87/0,00 (16,74)	0,88/0,00 (17,49)
ASW	1,00/- (5,63)	-/0,00 (19,99)	0,00/0,00 (1,01)	0,01/0,00 (2,02)	0,14/0,00 (1,56)	0,20/0,00 (2,92)	0,46/0,01 (4,72)	0,59/0,09 (6,08)
ASW × SW	1,00/- (15,43)	-/0,00 (19,92)	0,00/0,00 (1,00)	0,01/0,00 (2,02)	0,32/0,00 (2,91)	0,47/0,00 (5,67)	0,77/0,00 (10,92)	0,84/0,01 (13,40)
MSW	1,00/- (5,60)	-/0,01 (19,77)	0,00/0,00 (1,00)	0,00/0,00 (2,01)	0,55/0,00 (3,54)	0,36/0,00 (3,77)	0,38/0,02 (4,12)	0,46/0,11 (4,58)
MSW × SW	1,00/- (14,50)	-/0,01 (19,82)	0,00/0,00 (1,00)	0,00/0,00 (2,01)	0,60/0,00 (4,95)	0,53/0,00 (6,02)	0,71/0,00 (9,62)	0,80/0,02 (11,92)
cMCP	0,32/- (0,84)	-/0,00 (20,00)	0,57/0,00 (4,17)	0,45/0,00 (5,13)	0,67/0,00 (4,58)	0,56/0,00 (5,73)	0,42/0,04 (4,30)	0,30/0,24 (3,00)
gel	0,25/- (0,32)	-/0,04 (19,11)	0,00/0,00 (1,00)	0,00/0,00 (2,00)	0,02/0,07 (0,97)	0,00/0,59 (0,83)	0,00/0,63 (0,75)	0,02/0,65 (0,69)
SGL	0,40/- (2,12)	-/0,00 (20,00)	0,81/0,00 (8,31)	0,73/0,00 (9,34)	0,88/0,00 (10,56)	0,79/0,00 (11,05)	0,66/0,01 (8,10)	0,56/0,12 (6,40)
IPF-Lasso								
IPF-Lasso1	0,64/- (1,28)	-/0,00 (20,00)	0,46/0,00 (2,88)	0,25/0,00 (3,22)	0,60/0,00 (3,81)	0,37/0,00 (3,94)	0,23/0,01 (3,13)	0,19/0,10 (2,68)
IPF-Lasso2	0,64/- (1,28)	-/0,00 (20,00)	0,73/0,00 (6,92)	0,72/0,00 (8,78)	0,90/0,00 (10,82)	0,79/0,00 (10,88)	0,64/0,03 (7,28)	0,51/0,20 (5,03)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9)))

IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9)))

FIGURE B.1 : *Résultats de la simulation (a) au niveau des groupes de biomarqueurs. Boxplots du nombre des groupes de biomarqueurs sélectionnés dans tous les scénarios*



*Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. La boîte délimite l'intervalle interquartile (IQR) et contient une ligne horizontale correspondant à la médiane ; en dehors de la boîte, les moustaches de style Tukey s'étendent jusqu'à un maximum de 1,5*IQR au-delà de la boîte. Les losanges noirs représentent le nombre moyen de biomarqueurs sélectionnés. Les lignes noires en pointillés indiquent le nombre de biomarqueurs réellement actifs dans les scénarios alternatifs. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée*

TABLEAU B.4 : *Résultats de la simulation (b) au niveau des biomarqueurs. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)*

Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 ($l = 0$; $q = 0$)	**Scénario 2** ($l = 20$; $q = 100$)	**Scénario 3** ($l = 1$; $q = 8$)	**Scénario 4** ($l = 2$; $q = 8$)	**Scénario 5** ($l = 1$; $q = 32$)	**Scénario 6** ($l = 2$; $q = 32$)	**Scénario 7** ($l = 2$; $q = 16$)	**Scénario 8** ($l = 2$; $q = 8$)
		HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
Lasso standard	1,00/- (11,62)	0,48/0,13 (168,26)	0,81/0,04 (43,34)	0,81/0,05 (44,19)	0,71/0,49 (58,84)	0,72/0,50 (58,64)	0,76/0,53 (36,46)	0,82/0,57 (24,60)
Lasso adaptatif								
AC	1,00/- (14,95)	0,48/0,14 (165,09)	0,66/0,03 (25,66)	0,70/0,04 (28,31)	0,50/0,40 (40,89)	0,53/0,41 (42,74)	0,70/0,46 (33,76)	0,80/0,54 (26,63)
PCA	1,00/- (8,94)	0,57/0,50 (116,99)	0,70/0,08 (32,34)	0,72/0,10 (32,40)	0,59/0,51 (38,69)	0,57/0,54 (37,86)	0,70/0,61 (24,23)	0,80/0,69 (17,29)
Lasso + PCA	1,00/- (8,49)	0,50/0,17 (165,69)	0,83/0,14 (47,67)	0,75/0,10 (35,64)	0,74/0,58 (46,69)	0,65/0,54 (44,78)	0,70/0,56 (27,59)	0,75/0,61 (18,00)
SW	1,00/- (53,46)	0,14/0,15 (98,27)	0,29/0,03 (12,58)	0,38/0,03 (16,20)	0,58/0,44 (45,13)	0,58/0,45 (45,04)	0,80/0,43 (48,70)	0,90/0,45 (50,87)
ASW	1,00/- (16,70)	0,49/0,17 (161,20)	0,56/0,03 (18,60)	0,61/0,04 (20,83)	0,34/0,38 (31,15)	0,39/0,38 (33,54)	0,60/0,44 (26,52)	0,76/0,52 (23,04)
ASW × SW	1,00/- (49,70)	0,15/0,19 (95,63)	0,05/0,03 (8,26)	0,11/0,04 (8,78)	0,22/0,37 (28,30)	0,24/0,39 (28,37)	0,64/0,35 (35,19)	0,86/0,38 (42,95)
MSW	1,00/- (16,29)	0,52/0,24 (159,40)	0,56/0,03 (18,58)	0,61/0,04 (21,16)	0,38/0,38 (33,55)	0,45/0,40 (35,89)	0,61/0,45 (25,46)	0,73/0,54 (19,55)
MSW × SW	1,00/- (46,75)	0,18/0,21 (96,30)	0,06/0,03 (8,30)	0,11/0,03 (8,86)	0,26/0,38 (29,22)	0,32/0,39 (31,07)	0,63/0,35 (34,08)	0,84/0,37 (39,43)
cMCP	0,32/- (0,98)	0,10/0,69 (34,76)	0,40/0,58 (7,42)	0,44/0,59 (7,93)	0,34/0,76 (12,52)	0,33/0,77 (12,04)	0,37/0,79 (6,56)	0,38/0,80 (3,81)
gel	0,30/- (10,23)	0,03/0,58 (43,47)	0,10/0,06 (8,46)	0,12/0,06 (8,67)	0,31/0,45 (40,81)	0,14/0,80 (15,51)	0,12/0,84 (8,06)	0,10/0,82 (6,29)
SGL	0,39/- (9,34)	0,49/0,06 (186,25)	0,77/0,01 (45,82)	0,77/0,01 (47,08)	0,65/0,38 (67,17)	0,65/0,37 (66,59)	0,67/0,35 (43,63)	0,67/0,36 (30,03)
IPF-Lasso								
IPF-Lasso1	0,62/- (2,49)	0,37/0,16 (134,70)	0,44/0,04 (14,81)	0,46/0,04 (15,53)	0,28/0,44 (25,59)	0,30/0,46 (25,48)	0,40/0,49 (14,65)	0,49/0,56 (8,73)
IPF-Lasso2	0,37/- (1,84)	0,39/0,17 (136,80)	0,56/0,05 (19,60)	0,57/0,06 (19,68)	0,49/0,55 (30,73)	0,49/0,56 (29,70)	0,50/0,61 (14,71)	0,50/0,67 (7,92)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9)))

IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9)))

TABLEAU B.5 : *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Taux de fausses découvertes / taux de faux négatifs et (le nombre moyen de groupes de biomarqueurs sélectionnés)*

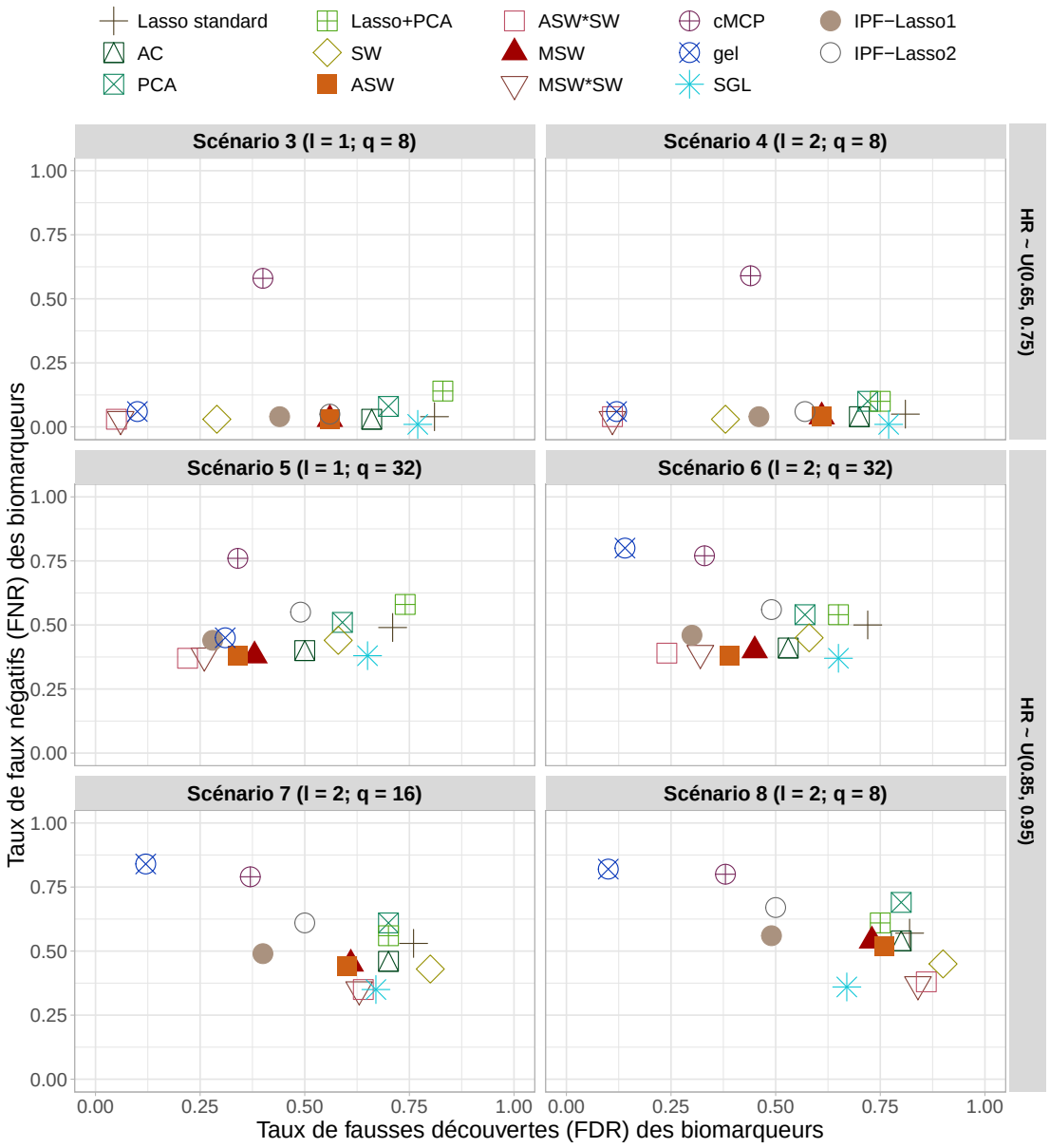
Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 ($l = 0; q = 0$)	**Scénario 2** ($l = 20; q = 100$)	**Scénario 3** ($l = 1; q = 8$)	**Scénario 4** ($l = 2; q = 8$)	**Scénario 5** ($l = 1; q = 32$)	**Scénario 6** ($l = 2; q = 32$)	**Scénario 7** ($l = 2; q = 16$)	**Scénario 8** ($l = 2; q = 8$)
		HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
Lasso standard	1,00/- (7,15)	-/0,00 (20,00)	0,93/0,00 (14,73)	0,87/0,00 (15,37)	0,94/0,00 (15,86)	0,87/0,00 (16,30)	0,84/0,00 (13,68)	0,80/0,04 (11,40)
Lasso adaptatif								
AC	1,00/- (6,28)	-/0,00 (20,00)	0,73/0,00 (5,50)	0,67/0,00 (7,33)	0,82/0,00 (7,24)	0,74/0,00 (8,84)	0,75/0,00 (9,32)	0,74/0,06 (9,05)
PCA	1,00/- (3,05)	-/0,19 (16,17)	0,58/0,03 (4,70)	0,51/0,05 (5,33)	0,73/0,12 (4,92)	0,64/0,17 (5,72)	0,63/0,26 (4,93)	0,66/0,38 (4,48)
$Lasso + PCA$	1,00/- (4,72)	-/0,00 (19,96)	0,85/0,08 (8,15)	0,66/0,05 (6,97)	0,91/0,17 (8,92)	0,80/0,10 (9,41)	0,73/0,10 (7,86)	0,69/0,17 (6,83)
SW	1,00/- (16,85)	-/0,00 (20,00)	0,50/0,00 (3,97)	0,45/0,00 (6,21)	0,91/0,00 (12,93)	0,84/0,00 (13,51)	0,87/0,00 (15,59)	0,87/0,00 (16,33)
ASW	1,00/- (5,42)	-/0,00 (19,99)	0,01/0,00 (1,01)	0,01/0,00 (2,03)	0,22/0,00 (1,75)	0,29/0,00 (3,34)	0,45/0,03 (4,65)	0,58/0,10 (5,89)
$ASW \times SW$	1,00/- (14,49)	-/0,00 (19,91)	0,00/0,00 (1,00)	0,00/0,00 (2,01)	0,47/0,00 (3,95)	0,43/0,00 (5,11)	0,73/0,00 (9,72)	0,82/0,02 (12,74)
MSW	1,00/- (4,93)	-/0,01 (19,76)	0,00/0,00 (1,00)	0,00/0,00 (2,00)	0,41/0,00 (2,56)	0,36/0,00 (3,74)	0,37/0,03 (3,94)	0,45/0,13 (4,27)
$MSW \times SW$	1,00/- (13,16)	-/0,01 (19,80)	0,00/0,00 (1,00)	0,00/0,00 (2,01)	0,59/0,00 (4,50)	0,55/0,00 (6,03)	0,71/0,00 (8,86)	0,79/0,03 (11,10)
cMCP	0,32/- (0,85)	-/0,00 (19,99)	0,56/0,00 (3,90)	0,47/0,00 (5,18)	0,68/0,00 (4,63)	0,53/0,00 (5,36)	0,41/0,03 (4,28)	0,34/0,22 (3,24)
gel	0,30/- (0,35)	-/0,08 (18,34)	0,00/0,00 (1,00)	0,00/0,00 (2,00)	0,01/0,34 (0,68)	0,00/0,58 (0,84)	0,01/0,58 (0,86)	0,03/0,62 (0,81)
SGL	0,39/- (2,05)	-/0,00 (20,00)	0,83/0,00 (8,33)	0,73/0,00 (9,17)	0,87/0,00 (9,94)	0,77/0,00 (10,47)	0,66/0,01 (8,05)	0,57/0,12 (6,39)
IPF-Lasso								
IPF-Lasso1	0,62/- (1,29)	-/0,00 (20,00)	0,45/0,00 (2,62)	0,25/0,00 (3,43)	0,55/0,00 (3,54)	0,35/0,00 (3,97)	0,22/0,01 (3,17)	0,22/0,10 (3,01)
IPF-Lasso2	0,37/- (1,33)	-/0,00 (20,00)	0,84/0,00 (7,99)	0,73/0,00 (8,76)	0,88/0,00 (9,70)	0,78/0,00 (10,30)	0,64/0,03 (7,07)	0,51/0,20 (4,93)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9)))

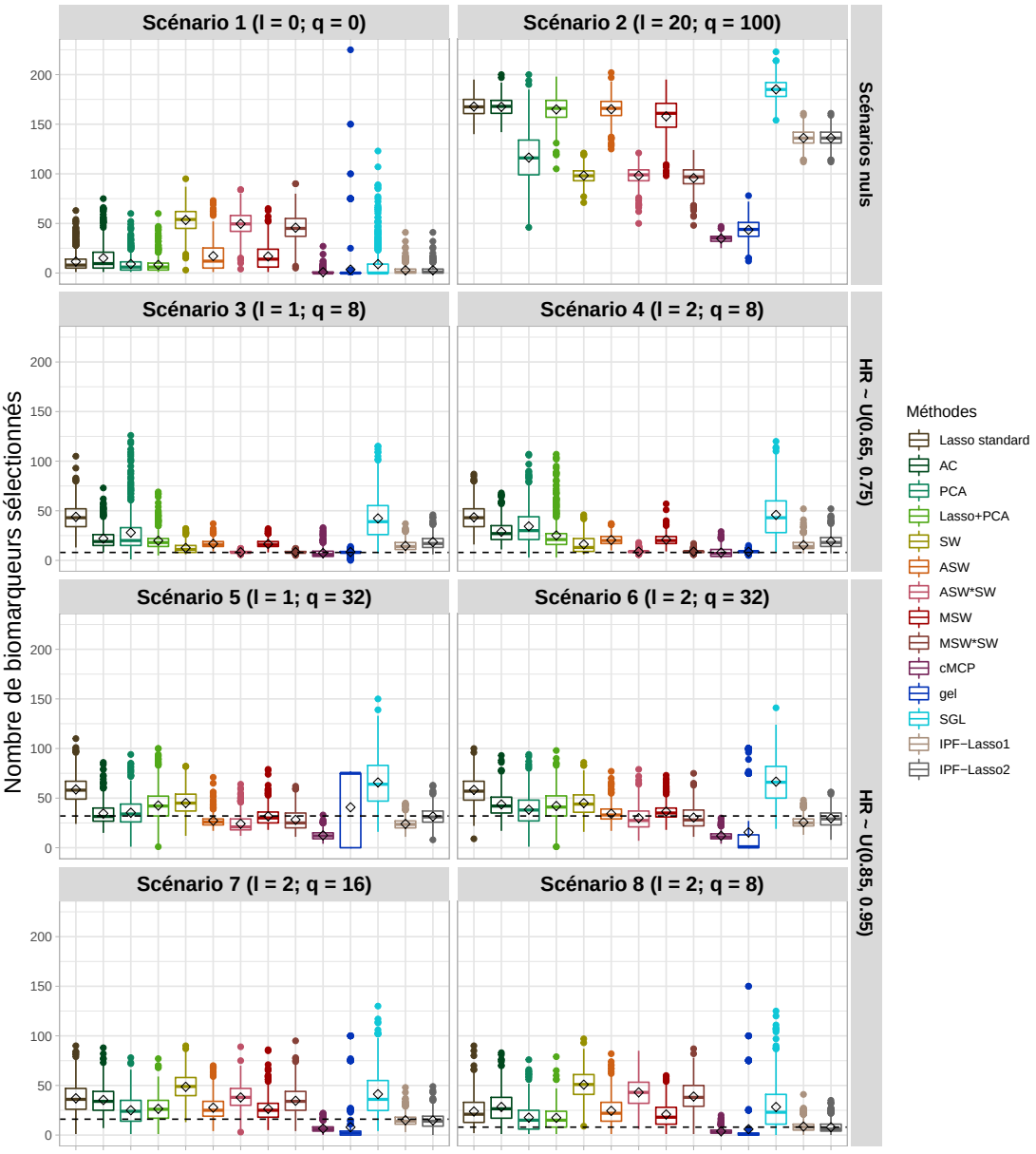
IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9)))

FIGURE B.2 : *Résultats de la simulation (b) au niveau des biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs*



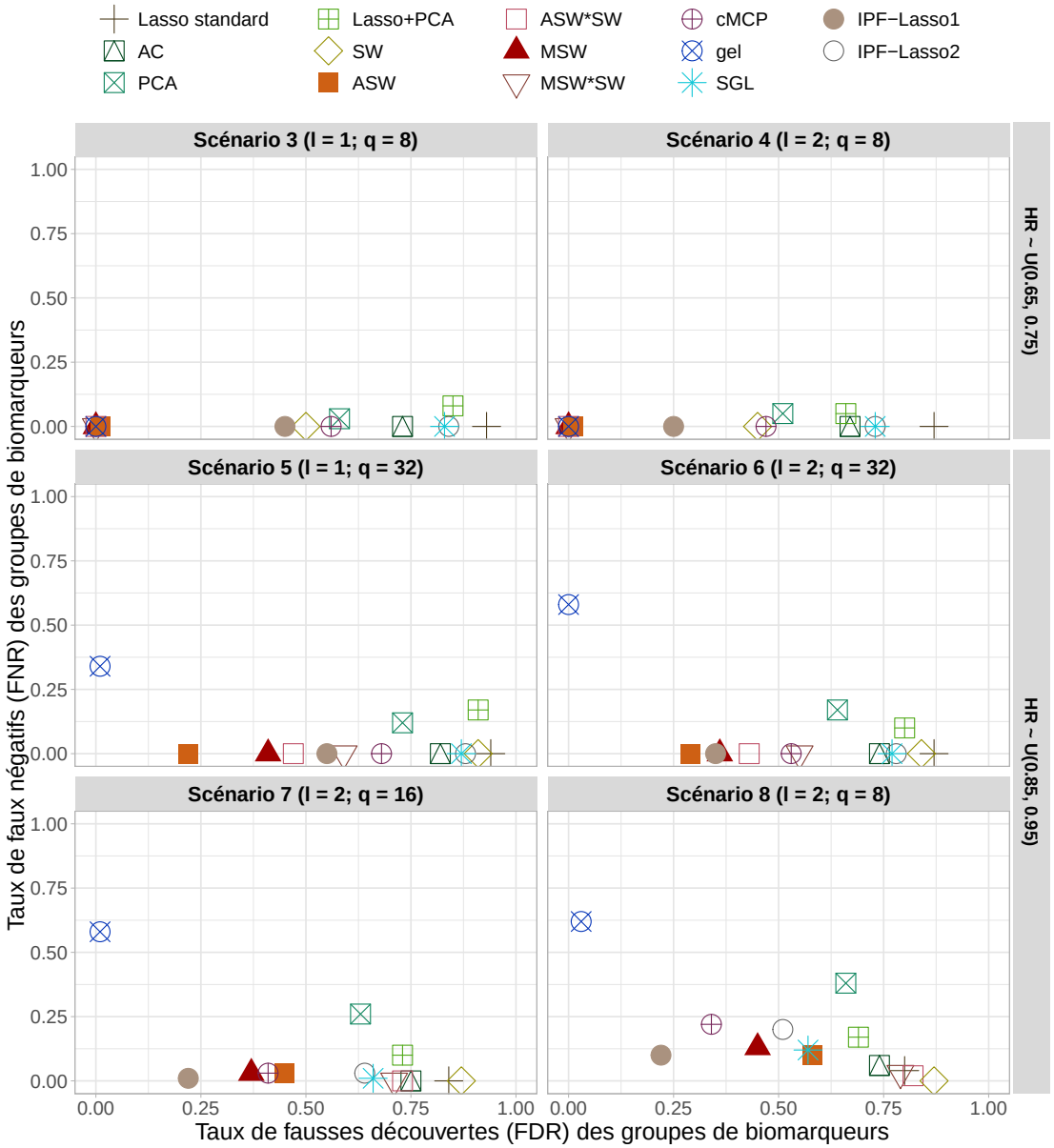
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.3 : *Résultats de la simulation (b) au niveau des biomarqueurs. Boxplots du nombre de biomarqueurs sélectionnés dans tous les scénarios*



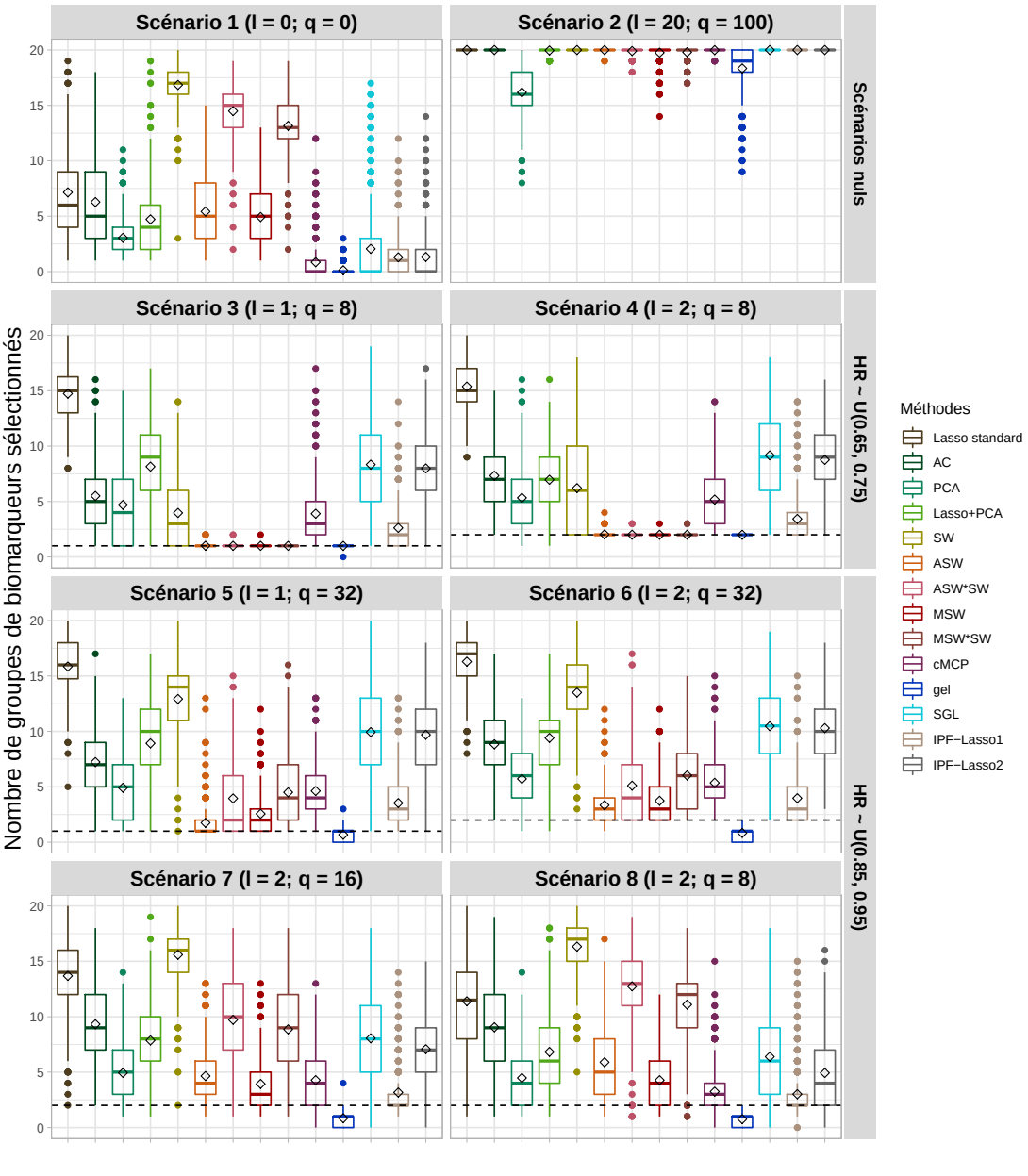
*Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. La boîte délimite l'intervalle interquartile (IQR) et contient une ligne horizontale correspondant à la médiane ; en dehors de la boîte, les moustaches de style Tukey s'étendent jusqu'à un maximum de 1,5*IQR au-delà de la boîte. Les losanges noirs représentent le nombre moyen de biomarqueurs sélectionnés. Les lignes noires en pointillés indiquent le nombre de biomarqueurs réellement actifs dans les scénarios alternatifs. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée*

FIGURE B.4 : *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Taux de faux négatifs (FNR) en fonction du taux de fausses découvertes (FDR) des biomarqueurs dans les scénarios alternatifs*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.5 : *Résultats de la simulation (b) au niveau des groupes de biomarqueurs. Boxplots du nombre des groupes de biomarqueurs sélectionnés dans tous les scénarios*



.

*Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. La boîte délimite l'intervalle interquartile (IQR) et contient une ligne horizontale correspondant à la médiane ; en dehors de la boîte, les moustaches de style Tukey s'étendent jusqu'à un maximum de 1,5*IQR au-delà de la boîte. Les losange noirs représentent le nombre moyen de biomarqueurs sélectionnés. Les lignes noires en pointillés indiquent le nombre de biomarqueurs réellement actifs dans les scénarios alternatifs. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée*

TABLEAU B.6 : *Résultats de la simulation (b). La moyenne et l'erreur standard empirique (ESE) de l'AUC à 5 ans*

Méthodes	**Scénario 3** ($l = 1$; $q = 8$)	**Scénario 4** ($l = 2$; $q = 8$)	**Scénario 5** ($l = 1$; $q = 32$)	**Scénario 6** ($l = 2$; $q = 32$)	**Scénario 7** ($l = 2$; $q = 16$)	**Scénario 8** ($l = 2$; $q = 8$)
	HR$\sim U(0,65,\ 0,75)$		HR$\sim U(0,85,\ 0,95)$			
Lasso standard	0,88 (0,03)	0,87 (0,03)	0,75 (0,04)	0,75 (0,04)	0,67 (0,05)	0,61 (0,05)
Lasso adaptatif						
AC	0,88 (0,03)	0,88 (0,02)	0,77 (0,04)	0,77 (0,04)	0,68 (0,04)	0,61 (0,05)
PCA	0,86 (0,07)	0,86 (0,05)	0,73 (0,10)	0,73 (0,07)	0,65 (0,07)	0,58 (0,07)
$Lasso + PCA$	0,83 (0,11)	0,86 (0,04)	0,70 (0,10)	0,73 (0,06)	0,67 (0,05)	0,61 (0,06)
SW	0,88 (0,03)	0,88 (0,03)	0,75 (0,04)	0,75 (0,04)	0,66 (0,05)	0,59 (0,05)
ASW	0,88 (0,02)	0,88 (0,03)	0,78 (0,03)	0,77 (0,04)	0,69 (0,04)	0,61 (0,05)
$ASW \times SW$	0,89 (0,02)	0,88 (0,02)	0,77 (0,04)	0,77 (0,04)	0,67 (0,05)	0,59 (0,05)
MSW	0,88 (0,02)	0,88 (0,02)	0,78 (0,04)	0,77 (0,04)	0,69 (0,05)	0,62 (0,05)
$MSW \times SW$	0,89 (0,02)	0,88 (0,02)	0,77 (0,04)	0,76 (0,04)	0,67 (0,05)	0,59 (0,05)
cMCP	0,87 (0,03)	0,86 (0,03)	0,75 (0,04)	0,74 (0,04)	0,67 (0,05)	0,61 (0,06)
gel	0,88 (0,03)	0,88 (0,02)	0,65 (0,11)	0,61 (0,09)	0,60 (0,07)	0,58 (0,07)
SGL	0,88 (0,02)	0,88 (0,02)	0,77 (0,04)	0,76 (0,04)	0,69 (0,05)	0,62 (0,06)
IPF-Lasso						
IPF-Lasso1	0,88 (0,02)	0,88 (0,02)	0,78 (0,04)	0,77 (0,04)	0,70 (0,04)	0,63 (0,05)
IPF-Lasso2	0,88 (0,03)	0,88 (0,02)	0,76 (0,04)	0,75 (0,04)	0,68 (0,05)	0,62 (0,06)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and 10 de taille 75. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

IPF-Lasso1 : pflist1 = list(c(1,rep(2,19)), c(2,1,rep(2,18)), c(rep(2,10),1,rep(2,9)), c(1,1,rep(2,18)), c(1,rep(2,9),1,rep(2,9)), c(2,1,rep(2,8),1,rep(2,9)), c(1,1,rep(2,8),1,rep(2,9)))

IPF-Lasso2 : pflist2 = list(c(2,rep(1,19)), c(1,2,rep(1,18)), c(rep(1,10),2,rep(1,9)), c(2,2,rep(1,18)), c(2,rep(1,9),2,rep(1,9)), c(1,2,rep(1,8),2,rep(1,9)), c(2,2,rep(1,8),2,rep(1,9)))

TABLEAU B.7 : *Résultats analyse de sensibilité au niveau des biomarqueurs pour le choix des paramètres de réglage des méthodes gel et SGL. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)*

	Scénarios nuls (sans effets de groupe)		**Scénarios alternatifs (avec effets de groupe)**					
Méthodes	**Scénario 1**	**Scénario 2**	**Scénario 3**	**Scénario 4**	**Scénario 5**	**Scénario 6**	**Scénario 7**	**Scénario 8**
	($l = 0; q = 0$)	**($l = 20; q = 100$)**	**($l = 1; q = 8$)**	**($l = 2; q = 8$)**	**($l = 1; q = 32$)**	**($l = 2; q = 32$)**	**($l = 2; q = 16$)**	**($l = 2; q = 8$)**
		HR~ $U(0,85,\ 0,95)$	HR~ $U(0,65,\ 0,75)$		HR~ $U(0,85,\ 0,95)$			
gel	0,81/- (3,82)	0,40/0,13 (141,03)	0,37/0,04 (16,95)	0,43/0,05 (19,60)	0,55/0,53 (36,54)	0,55/0,53 (36,11)	0,54/0,59 (18,23)	0,54/0,62 (10,64)
SGL	0,44/- (23,56)	0,49/0,05 (189,64)	0,79/0,01 (73,89)	0,77/0,01 (55,02)	0,67/0,06 (143,04)	0,75/0,09 (153,95)	0,75/0,19 (88,57)	0,69/0,37 (45,31)

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs de l'étude de simulation (a) et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs
Dans les scénarios 1, 2, 6, 7, et 8, la méthode gel avait 358, 44, 2, 3, et 53 itérations qui n'ont pas convergé, respectivement

TABLEAU B.8 : Résultats analyse de sensibilité au niveau des groupes de biomarqueurs pour le choix des paramètres de réglage des méthodes gel et SGL. Taux de fausses découvertes FDR / taux de faux négatifs FNR et (le nombre moyen de groupes de biomarqueurs sélectionnés)

Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 (l = **0**; q = **0**)	**Scénario 2** (l = **20**; q = **100**)	**Scénario 3** (l = **1**; q = **8**)	**Scénario 4** (l = **2**; q = **8**)	**Scénario 5** (l = **1**; q = **32**)	**Scénario 6** (l = **2**; q = **32**)	**Scénario 7** (l = **2**; q = **16**)	**Scénario 8** (l = **2**; q = **8**)
		HR$\sim U(0,85,\ 0,95)$	HR$\sim U(0,65,\ 0,75)$		HR$\sim U(0,85,\ 0,95)$			
gel	0,81/- (2,92)	-/0,00 (20,00)	0,32/0,00 (5,03)	0,41/0,00 (7,40)	0,90/0,00 (12,41)	0,82/0,00 (12,91)	0,68/0,02 (8,66)	0,55/0,11 (5,66)
SGL	0,44/- (2,11)	-/0,00 (20,00)	0,70/0,00 (5,96)	0,66/0,00 (7,90)	0,65/0,00 (4,97)	0,63/0,00 (7,02)	0,58/0,01 (6,54)	0,53/0,14 (5,74)

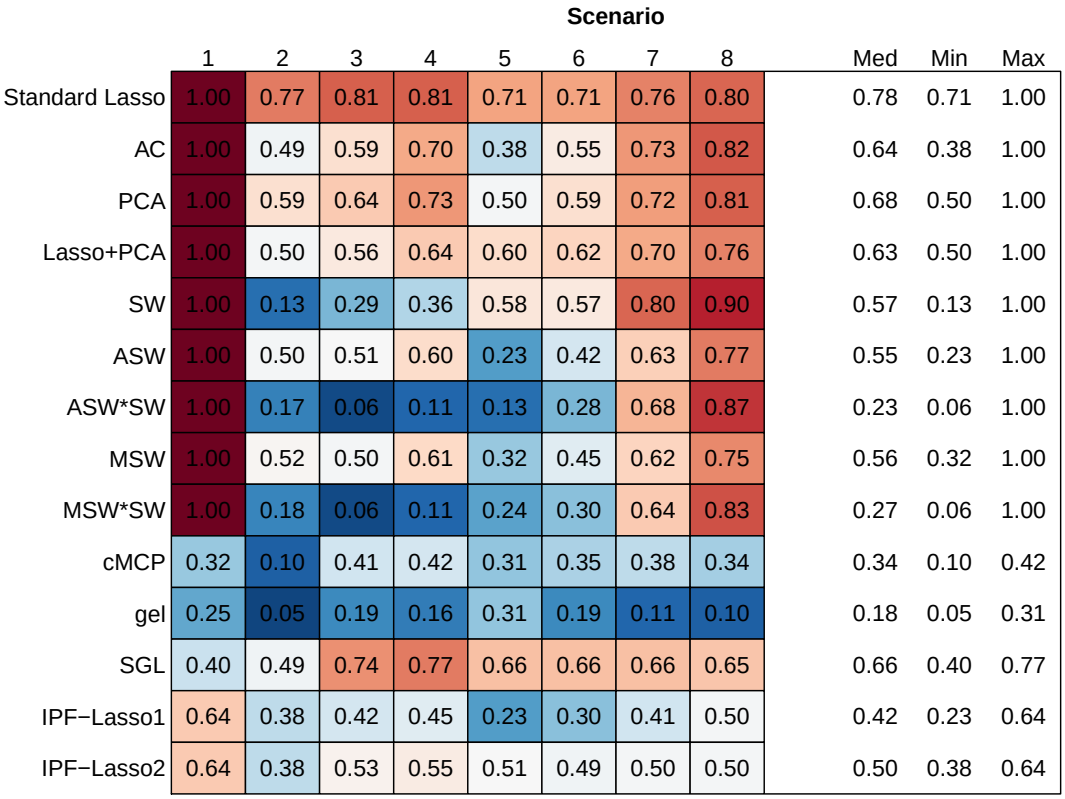
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs de l'étude de simulation (a) et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs
Dans les scénarios 1, 2, 6, 7, et 8, la méthode gel avait 358, 44, 2, 3, et 53 itérations qui n'ont pas convergé, respectivement

TABLEAU B.9 : *Résultats de l'étude de sensibilité avec $n = 50$ patients de la simulation (a) au niveau des biomarqueurs. Taux de fausses découvertes FDR / Taux de faux négatifs FNR et (le nombre moyen de biomarqueurs sélectionnés)*

Méthodes	**Scénarios nuls (sans effets de groupe)**		**Scénarios alternatifs (avec effets de groupe)**					
	Scénario 1 ($l = 0; q = 0$)	**Scénario 2** ($l = 20; q = 100$)	**Scénario 3** ($l = 1; q = 8$)	**Scénario 4** ($l = 2; q = 8$)	**Scénario 5** ($l = 1; q = 32$)	**Scénario 6** ($l = 2; q = 32$)	**Scénario 7** ($l = 2; q = 16$)	**Scénario 8** ($l = 2; q = 8$)
		HR$\sim U(0,85,\ 0,95)$	HR$\sim U(0,65,\ 0,75)$		HR$\sim U(0,85,\ 0,95)$			
Lasso standard	1,00/-(7,55)	0,59/0,96(11,06)	0,76/0,66(12,97)	0,76/0,67(12,79)	0,84/0,96(9,26)	0,83/0,96(8,33)	0,91/0,96(8,08)	0,95/0,96(7,91)
Lasso adaptatif								
AC	1,00/- (8,69)	0,64/0,96 (12,48)	0,71/0,56 (13,26)	0,77/0,64 (14,09)	0,74/0,91 (11,17)	0,81/0,94 (10,58)	0,91/0,95 (9,50)	0,95/0,96 (9,68)
PCA	1,00/- (6,99)	0,70/0,97 (9,12)	0,78/0,71 (10,15)	0,82/0,80 (9,67)	0,82/0,94 (8,11)	0,85/0,96 (7,86)	0,93/0,97 (7,12)	0,97/0,97 (7,48)
$Lasso + PCA$	1,00/- (5,67)	0,65/0,97 (8,21)	0,64/0,58 (10,40)	0,75/0,75 (9,27)	0,75/0,93 (7,42)	0,81/0,97 (6,41)	0,91/0,97 (6,16)	0,96/0,97 (6,17)
ASW	0,99/- (9,58)	0,66/0,96 (12,87)	0,65/0,52 (12,50)	0,74/0,63 (13,16)	0,69/0,89 (11,66)	0,79/0,94 (10,87)	0,91/0,95 (10,31)	0,96/0,96 (10,03)
$ASW \times SW$	0,96/- (11,46)	0,62/0,94 (14,96)	0,57/0,56 (9,97)	0,67/0,61 (11,38)	0,76/0,92 (12,19)	0,81/0,93 (12,34)	0,89/0,94 (11,96)	0,93/0,95 (11,25)
MSW	0,99/- (9,90)	0,65/0,96 (12,33)	0,64/0,52 (11,83)	0,71/0,63 (12,00)	0,77/0,92 (11,24)	0,82/0,94 (10,95)	0,91/0,95 (10,07)	0,96/0,96 (9,77)
$MSW \times SW$	0,97/- (10,91)	0,60/0,95 (13,44)	0,56/0,55 (9,84)	0,63/0,60 (10,38)	0,78/0,93 (11,31)	0,83/0,94 (11,67)	0,90/0,95 (10,86)	0,93/0,95 (10,63)

Résultats avec $n = 50$ patients, $p = 1000$ biomarqueurs de l'étude de simulation (a) et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. l : nombre de groupes actifs, q : nombre de biomarqueurs actifs, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.6 : *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.7 : *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR*

Scenario

	2	3	4	5	6	7	8	Med	Min	Max
Standard Lasso	0.46	0.04	0.04	0.49	0.50	0.54	0.57	0.49	0.04	0.57
AC	0.15	0.02	0.04	0.38	0.42	0.47	0.53	0.38	0.02	0.53
PCA	0.53	0.05	0.11	0.46	0.54	0.62	0.70	0.53	0.05	0.70
Lasso+PCA	0.18	0.03	0.09	0.49	0.54	0.57	0.61	0.49	0.03	0.61
SW	0.15	0.02	0.03	0.45	0.45	0.44	0.44	0.44	0.02	0.45
ASW	0.18	0.02	0.03	0.36	0.40	0.45	0.52	0.36	0.02	0.52
ASW*SW	0.19	0.02	0.03	0.39	0.39	0.36	0.38	0.36	0.02	0.39
MSW	0.24	0.02	0.03	0.37	0.40	0.45	0.53	0.37	0.02	0.53
MSW*SW	0.21	0.02	0.03	0.38	0.39	0.35	0.36	0.35	0.02	0.39
cMCP	0.69	0.58	0.60	0.75	0.77	0.79	0.81	0.75	0.58	0.81
gel	0.52	0.04	0.05	0.13	0.80	0.86	0.85	0.52	0.04	0.86
SGL	0.05	0.01	0.01	0.39	0.36	0.35	0.34	0.34	0.01	0.39
IPF−Lasso1	0.16	0.03	0.04	0.44	0.46	0.50	0.56	0.44	0.03	0.56
IPF−Lasso2	0.16	0.04	0.05	0.54	0.55	0.61	0.67	0.54	0.04	0.67

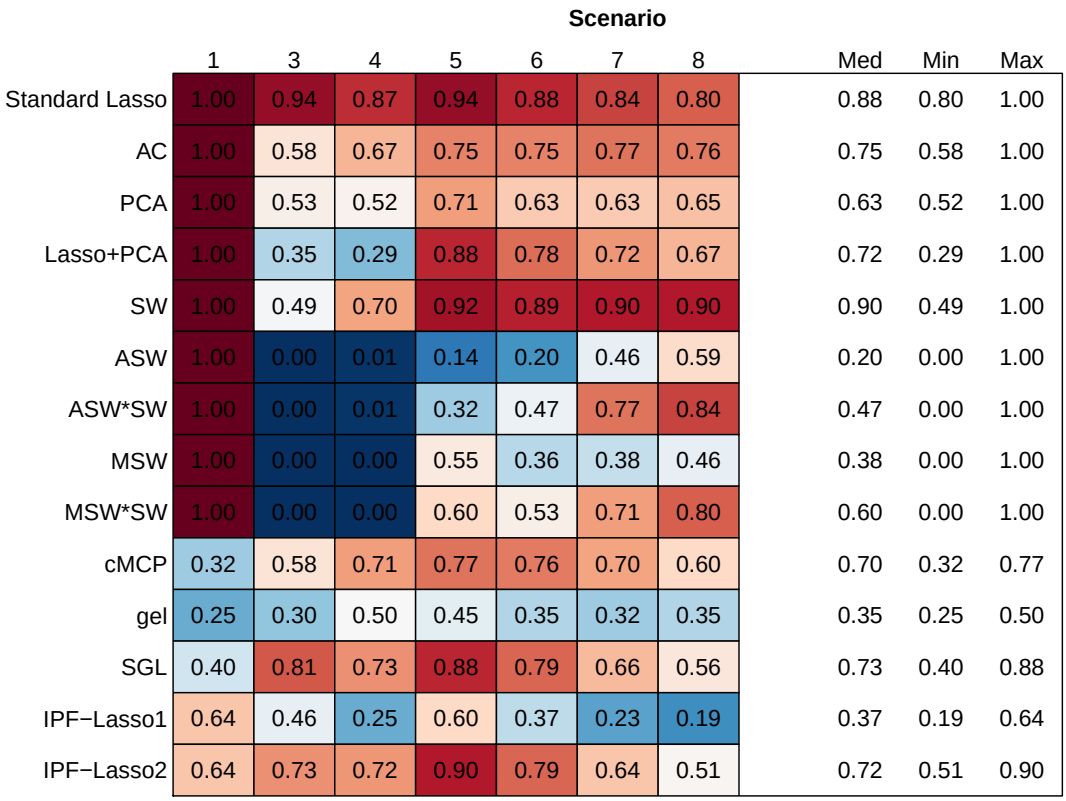
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.8 : *Résultats de la simulations (a) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores* $F1$

	Scenario									
	2	3	4	5	6	7	8	Med	Min	Max
Standard Lasso	0.29	0.32	0.31	0.36	0.36	0.29	0.24	0.31	0.24	0.36
AC	0.64	0.57	0.45	0.61	0.50	0.34	0.23	0.50	0.23	0.64
PCA	0.43	0.51	0.39	0.51	0.41	0.29	0.20	0.41	0.20	0.51
Lasso+PCA	0.62	0.59	0.49	0.44	0.40	0.33	0.26	0.44	0.26	0.62
SW	0.85	0.80	0.73	0.46	0.47	0.29	0.17	0.47	0.17	0.85
ASW	0.62	0.65	0.56	0.69	0.58	0.42	0.27	0.58	0.27	0.69
ASW*SW	0.82	0.96	0.92	0.70	0.64	0.40	0.21	0.70	0.21	0.96
MSW	0.58	0.65	0.55	0.65	0.57	0.43	0.30	0.57	0.30	0.65
MSW*SW	0.80	0.96	0.93	0.67	0.63	0.44	0.25	0.67	0.25	0.96
cMCP	0.46	0.46	0.43	0.36	0.33	0.29	0.26	0.36	0.26	0.46
gel	0.63	0.87	0.89	0.68	0.16	0.14	0.18	0.63	0.14	0.89
SGL	0.66	0.39	0.36	0.42	0.43	0.41	0.36	0.41	0.36	0.66
IPF−Lasso1	0.71	0.71	0.68	0.64	0.60	0.53	0.43	0.64	0.43	0.71
IPF−Lasso2	0.71	0.62	0.60	0.46	0.47	0.41	0.33	0.47	0.33	0.71

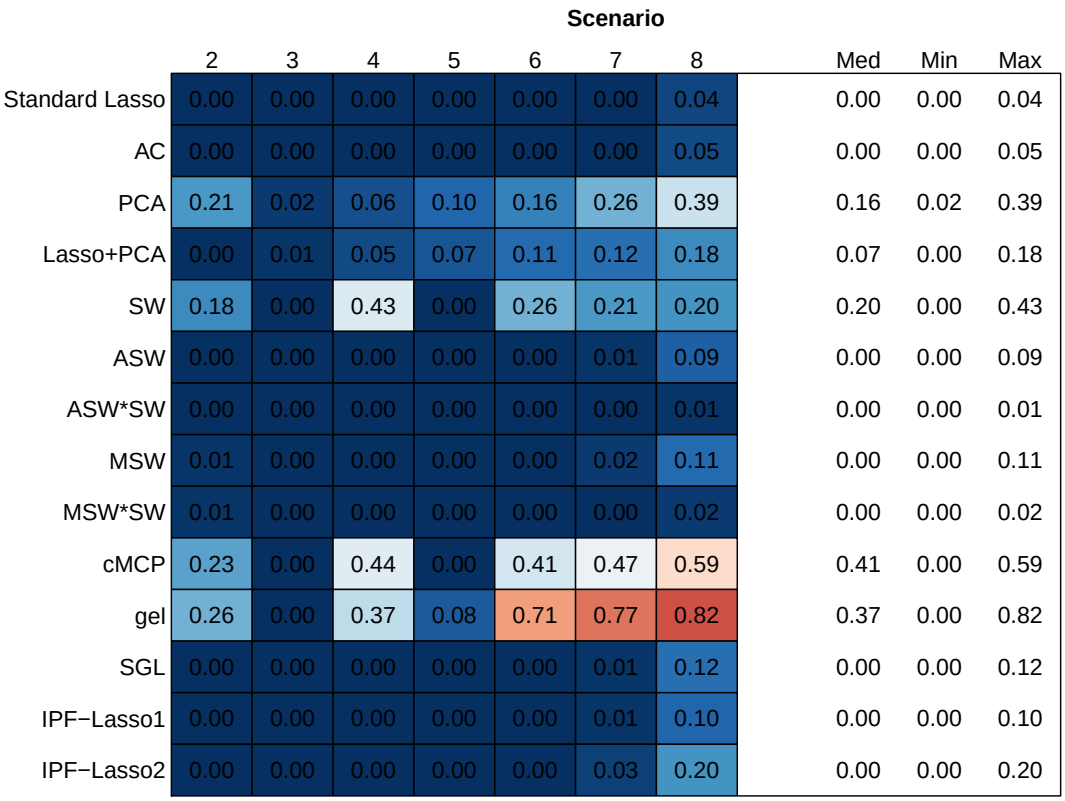
Résultats avec $n = 500$ *patients,* $p = 1000$ *biomarqueurs et* $R = 20$ *groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum,* AC *: moyenne des coefficients,* PCA *: analyse en composantes principales,* SW *: statistique de Wald univariée,* ASW *: moyenne de Wald univariée,* MSW *: maximum de Wald univariée*

FIGURE B.9 : *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR*



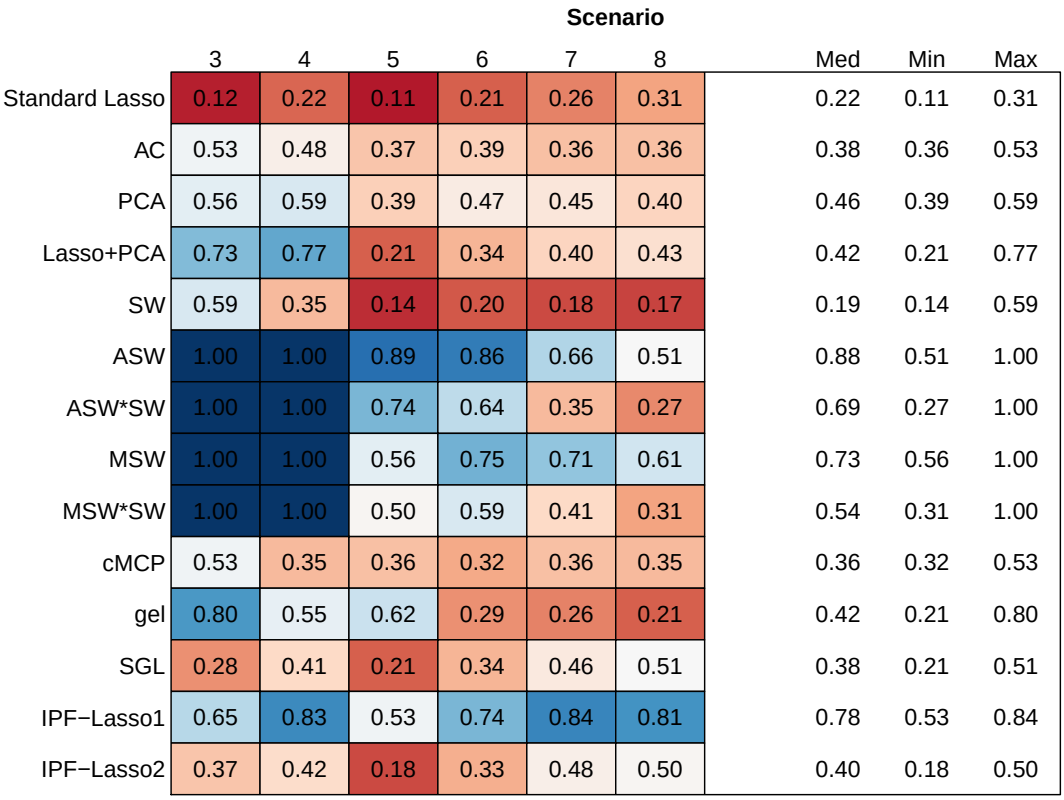
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.10 : *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.11 : *Résultats de la simulations (a) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs de taille égale. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.12 : *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR*

	Scenario										
	1	2	3	4	5	6	7	8	Med	Min	Max
Standard Lasso	1.00	0.48	0.81	0.81	0.71	0.72	0.76	0.82	0.78	0.48	1.00
AC	1.00	0.48	0.66	0.70	0.50	0.53	0.70	0.80	0.68	0.48	1.00
PCA	1.00	0.57	0.70	0.72	0.59	0.57	0.70	0.80	0.70	0.57	1.00
Lasso+PCA	1.00	0.50	0.83	0.75	0.74	0.65	0.70	0.75	0.74	0.50	1.00
SW	1.00	0.14	0.29	0.38	0.58	0.58	0.80	0.90	0.58	0.14	1.00
ASW	1.00	0.49	0.56	0.61	0.34	0.39	0.60	0.76	0.58	0.34	1.00
ASW*SW	1.00	0.15	0.05	0.11	0.22	0.24	0.64	0.86	0.23	0.05	1.00
MSW	1.00	0.52	0.56	0.61	0.38	0.45	0.61	0.73	0.58	0.38	1.00
MSW*SW	1.00	0.18	0.06	0.11	0.26	0.32	0.63	0.84	0.29	0.06	1.00
cMCP	0.32	0.10	0.40	0.44	0.34	0.33	0.37	0.38	0.36	0.10	0.44
gel	0.30	0.03	0.10	0.12	0.31	0.14	0.12	0.10	0.12	0.03	0.31
SGL	0.39	0.49	0.77	0.77	0.65	0.65	0.67	0.67	0.66	0.39	0.77
IPF−Lasso1	0.62	0.37	0.44	0.46	0.28	0.30	0.40	0.49	0.42	0.28	0.62
IPF−Lasso2	0.37	0.39	0.56	0.57	0.49	0.49	0.50	0.50	0.50	0.37	0.57

Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

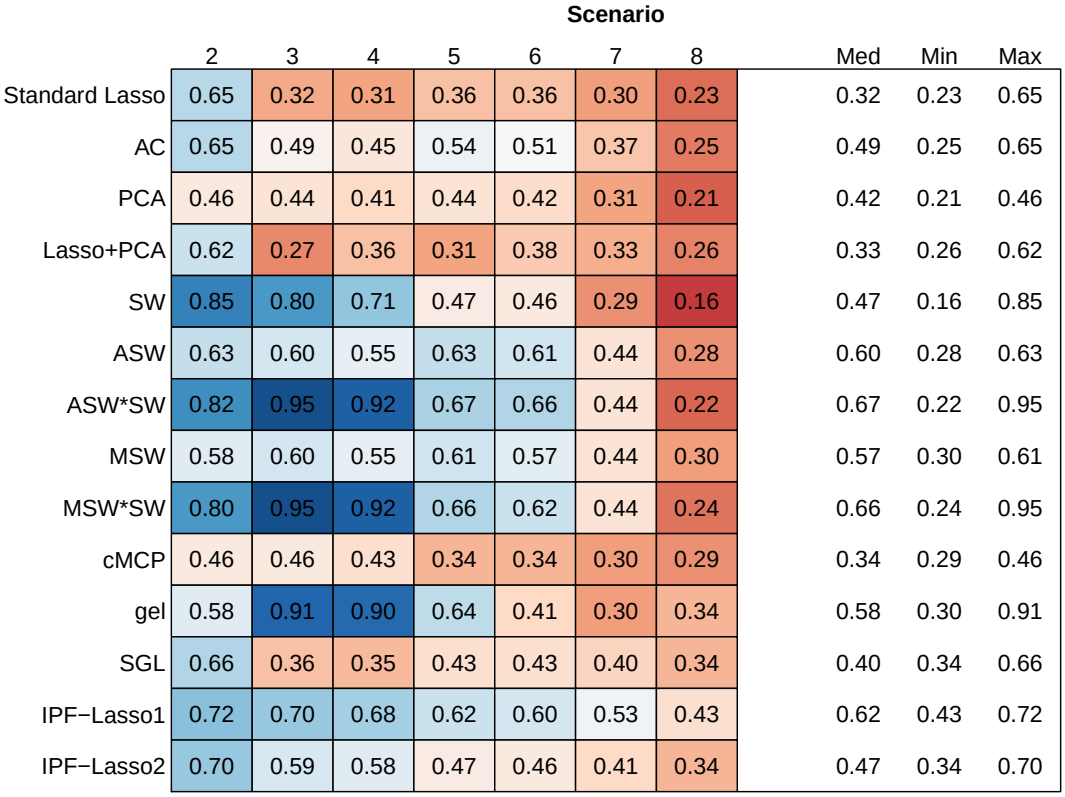
FIGURE B.13 : *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR*

Scenario

	2	3	4	5	6	7	8	Med	Min	Max
Standard Lasso	0.13	0.04	0.05	0.49	0.50	0.53	0.57	0.49	0.04	0.57
AC	0.14	0.03	0.04	0.40	0.41	0.46	0.54	0.40	0.03	0.54
PCA	0.50	0.08	0.10	0.51	0.54	0.61	0.69	0.51	0.08	0.69
Lasso+PCA	0.17	0.14	0.10	0.58	0.54	0.56	0.61	0.54	0.10	0.61
SW	0.15	0.03	0.03	0.44	0.45	0.43	0.45	0.43	0.03	0.45
ASW	0.17	0.03	0.04	0.38	0.38	0.44	0.52	0.38	0.03	0.52
ASW*SW	0.19	0.03	0.04	0.37	0.39	0.35	0.38	0.35	0.03	0.39
MSW	0.24	0.03	0.04	0.38	0.40	0.45	0.54	0.38	0.03	0.54
MSW*SW	0.21	0.03	0.03	0.38	0.39	0.35	0.37	0.35	0.03	0.39
cMCP	0.69	0.58	0.59	0.76	0.77	0.79	0.80	0.76	0.58	0.80
gel	0.58	0.06	0.06	0.45	0.80	0.84	0.82	0.58	0.06	0.84
SGL	0.06	0.01	0.01	0.38	0.37	0.35	0.36	0.35	0.01	0.38
IPF−Lasso1	0.16	0.04	0.04	0.44	0.46	0.49	0.56	0.44	0.04	0.56
IPF−Lasso2	0.17	0.05	0.06	0.55	0.56	0.61	0.67	0.55	0.05	0.67

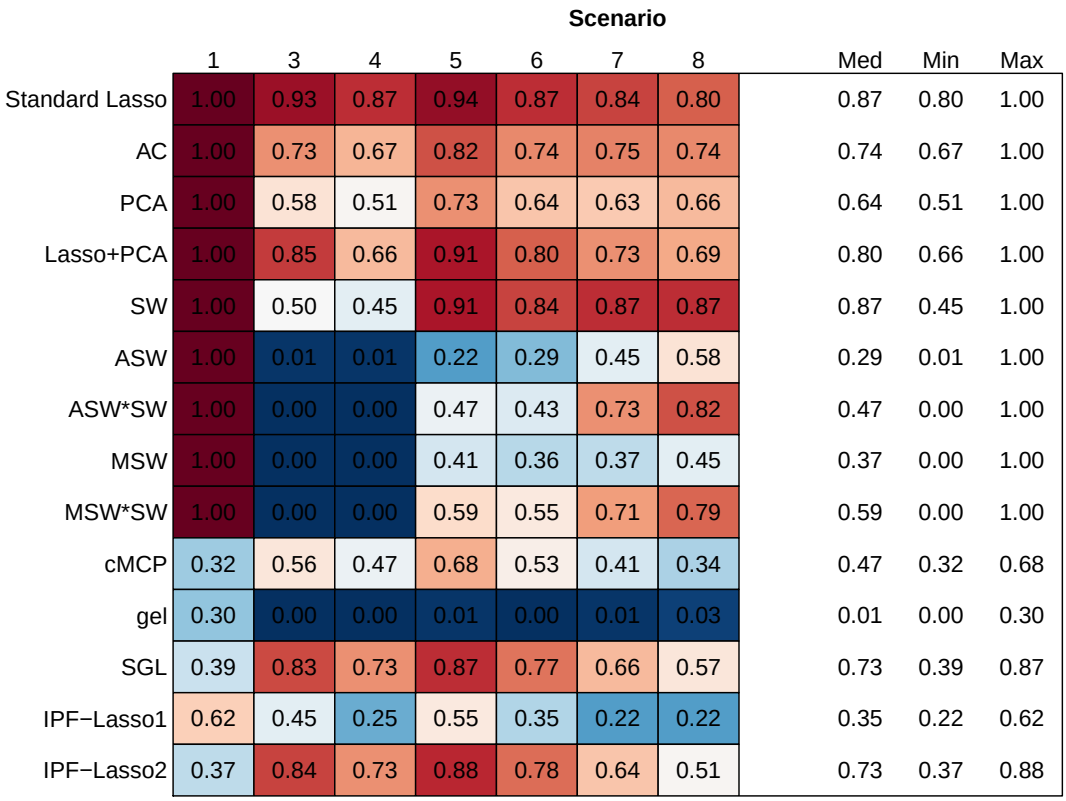
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.14 : *Résultats de la simulations (b) au niveau des biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores* $F1$



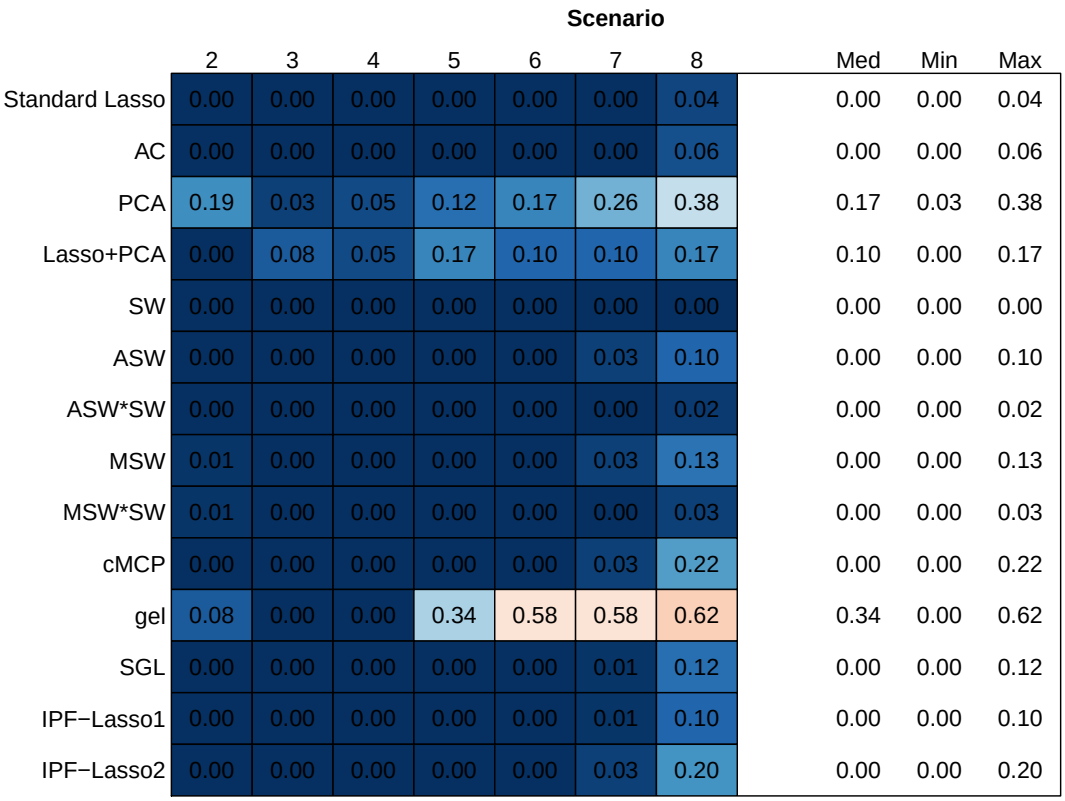
Résultats avec $n = 500$ *patients,* $p = 1000$ *biomarqueurs et* $R = 20$ *groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum,* AC *: moyenne des coefficients,* PCA *: analyse en composantes principales,* SW *: statistique de Wald univariée,* ASW *: moyenne de Wald univariée,* MSW *: maximum de Wald univariée*

FIGURE B.15 : *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de fausses découvertes FDR*



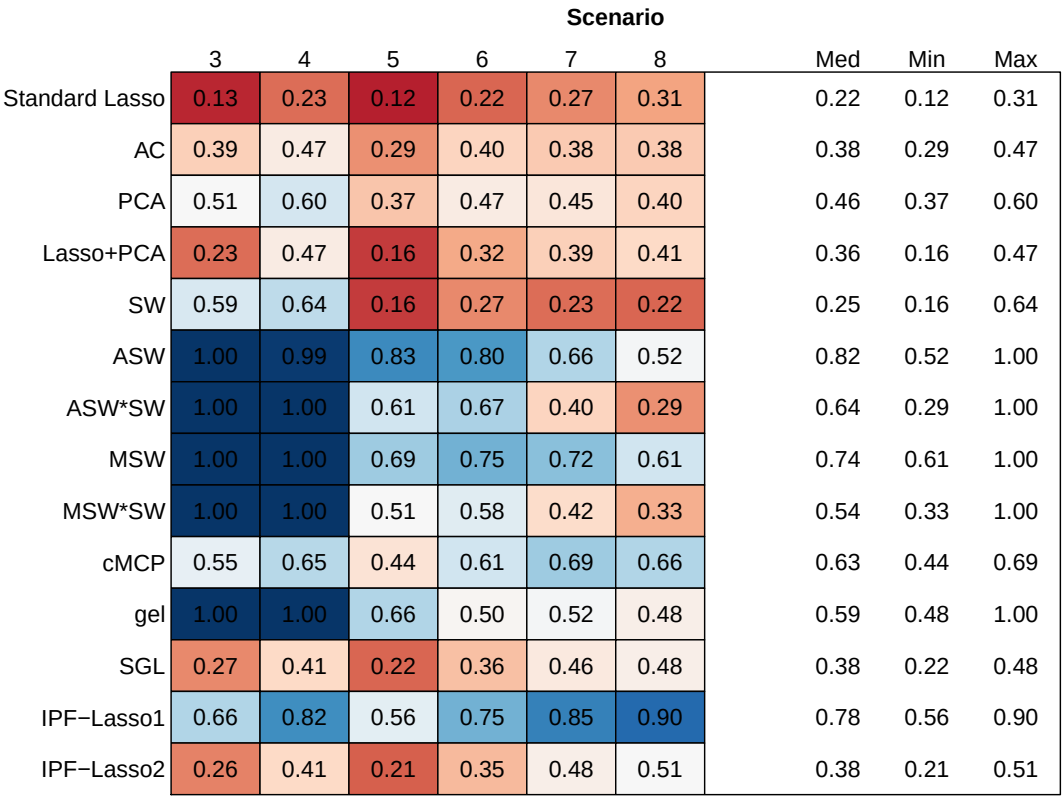
Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.16 : *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des taux de faux négatifs FNR*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

FIGURE B.17 : *Résultats de la simulations (b) au niveau des groupes de biomarqueurs. Heatmap et (médiane, minimum, maximum) des scores $F1$*



Résultats avec $n = 500$ patients, $p = 1000$ biomarqueurs et $R = 20$ groupes de biomarqueurs : 10 de taille 25 and de taille 75. Quantités moyennes sur 500 réplications. Med : Médiane, Min : Minimum, Max : Maximum, AC : moyenne des coefficients, PCA : analyse en composantes principales, SW : statistique de Wald univariée, ASW : moyenne de Wald univariée, MSW : maximum de Wald univariée

université PARIS-SACLAY

ÉCOLE DOCTORALE
Santé Publique
(EDSP)

Titre : Méthodes statistiques innovantes pour identifier des biomarqueurs dans les essais cliniques en oncologie

Mots clés : Sélection de biomarqueurs, Validation des critères de substitution, Analyse de survie, Sélection de variable en grande dimension, Régression pénalisée, Médecine de précision

Résumé : Les biomarqueurs ont un rôle de plus en plus important dans la recherche clinique en oncologie en raison de leur grande utilité. Ils ont le potentiel de remplacer un critère de substitution, de prédire le pronostic des patients, de guider les décisions thérapeutiques ou de sélectionner les patients éligibles pour un traitement. D'autre part, les exigences statistiques pour la validation d'un critère de substitution et la sélection de biomarqueurs pronostiques et prédictifs soulèvent de nombreux défis méthodologiques. L'objectif de cette thèse est de proposer de nouvelles approches pour l'identification de ces différentes classes de biomarqueurs. Des mesures de distance basées sur l'erreur de prédiction ont été proposées pour la validation d'un critère de substitution de type survie au niveau de l'étude. Elles permettent de mesurer la force de la corrélation entre les deux critères de jugement en termes d'erreur de prédiction. Afin de prendre en compte lappartenance des biomarqueurs pronostiques à des groupes préspécifiés, diverses stratégies de pondération ont été proposées pour la méthode lasso adaptatif. Ces pondérations sont basées sur le coefficient de régression de Cox univariée et la statistique de Wald univariée et elles cherchent à effectuer la sélection à deux niveaux : des groupes les plus pronostiques et des biomarqueurs pertinents au sein des groupes sélectionnés tout en éliminant les faux positifs. Pour les biomarqueurs prédictifs, des pondérations pour le lasso adaptatif basées sur le test de rapport de vraisemblance partielle ont été développées en vue de favoriser la contrainte hiérarchique de linteraction entre les biomarqueurs et le traitement. Des études de simulations impliquant plusieurs scénarios ont été menées pour évaluer les différentes approches présentées. Ces méthodes ont également été illustrées sur à travers des données réelles sur le cancer digestif avancé et le cancer du sein précoce.

Title : Innovative statistical methods for the identification of biomarkers in oncology

Keywords : Biomarker selection, Surrogate endpoint validation, Survival analysis, High dimensional variable selection, Penalized regression, Precision medicine

Abstract : Biomarkers have great potential for use in clinical oncology research. They can substitute a surrogate endpoint, predict a future disease-related patient outcome and guide therapeutic decisions or select patients who are likely to benefit from a particular treatment. Nevertheless, the validation of a surrogate endpoint and the selection of prognostic and predictive biomarkers presents major methodological challenges. The objective of this thesis is to propose new approaches for the identification of these different classes of biomarkers. Distance measures based on prediction error have been proposed for the validation of a trial-level surrogate endpoint. It measures the strength of the correlation between the two criteria in terms of prediction error. In order to take into account the group structure of the biomarkers, different weighting strategies have been proposed for the adaptive lasso method. These weights are based on the univariate Cox regression coefficient and the univariate Wald statistic. They aim to perform bilevel variable selection : selecting prognostic groups as well as identifying relevant biomarkers of these groups. For predictive biomarkers, adaptive lasso weights based on the partial likelihood ratio test were developed to favor the hierarchical constraint of treatment-biomarker interactions. The different approaches presented were evaluated through simulation studies involving several scenarios. In addition, these methods were illustrated on different data sets on advanced gastric cancer and early breast cancer.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France