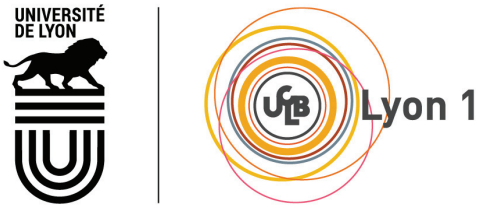
Méthodologie de l'évaluation des biomarqueurs prédictifs quantitatifs et de la détermination d'un seuil pour leur utilisation en médecine personnalisée

Yoann Blangero



N° d'ordre NNT : 2019LYSE1125

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED341
Evolution Ecosystèmes Microbiologie Modélisation

Spécialité de doctorat : Biostatistiques

Soutenue publiquement le 13/09/2019, par :
Yoann Blangero

Méthodologie de l'évaluation des biomarqueurs prédictifs quantitatifs et de la détermination d'un seuil pour leur utilisation en médecine personnalisée

---

Devant le jury composé de :

MAUCORT-BOULCH Delphine, PU-PH, Université Claude Bernard Lyon 1	Présidente
BINQUET Christine, PU-PH, Université de Bourgogne	Rapporteure
JACQMIN-GADDA Hélène, Directrice de recherche, Université de Bordeaux	Examinatrice
PAOLETTI Xavier, PhD, Institut Gustave Roussy	Rapporteur
SUBTIL Fabien, MCU-PH, Université Claude Bernard Lyon 1	Directeur de thèse
RABILLOUD Muriel, MCU-PH, Université Claude Bernard Lyon 1	Co-directrice de thèse

Résumé en français et en anglais

Résumé

En France, la recherche contre le cancer est un enjeu majeur de santé publique. On estime notamment que le nombre de nouveaux cas de cancer a plus que doublé entre 1980 et 2012. L'hétérogénéité des caractéristiques tumorales, pour un même cancer, impose des défis complexes dans la recherche de traitements efficaces. Dans ce contexte, des espoirs importants sont placés dans la recherche de biomarqueurs prédictifs reflétant les caractéristiques des patients ainsi que de leur tumeur afin d'orienter le choix de la stratégie thérapeutique. Par exemple, pour les cancers colorectaux métastatiques, il est maintenant reconnu que l'ajout de cetuximab (un anti-EGFR) à la chimiothérapie classique (ici le FOLFOX4), n'apporte un bénéfice qu'aux patients dont le gène *KRAS* est non muté. Le gène *KRAS* est ici un biomarqueur prédictif binaire, mais de nombreux biomarqueurs sont le résultat d'une quantification ou d'un dosage.

L'objectif de cette thèse est dans un premier temps, de quantifier la capacité globale d'un biomarqueur quantitatif à guider le choix du traitement. Après une revue de la littérature, une nouvelle méthode basée sur une extension des courbes ROC est proposée, et comparée aux méthodes existantes. Son principal avantage est d'être non paramétrique, et d'être indépendante de l'efficacité moyenne des traitements.

Dans un second temps, lorsqu'un biomarqueur prédictif quantitatif est étudié, la définition d'un seuil de marqueur au-delà duquel la première option de traitement sera préférée, et endeçà duquel la deuxième option de traitement sera préférée se pose. Une approche reposant sur la définition d'une fonction d'utilité est proposée permettant alors de tenir compte de l'efficacité des traitements ainsi que de leur impact sur la qualité de vie des patients. Une méthode Bayésienne d'estimation de ce seuil optimal est proposée.

Mots-clés : évaluation de biomarqueurs ; biomarqueurs prédictifs ; courbes ROC ; seuil optimal ; fonction d'utilité ; cancer colorectal

Title

Treatment selection markers in precision medicine: methodology of use and estimation of marker threshold

Abstract

In France, the cancer research is a major public health issue. The number of new cancer cases

nearly doubled between 1980 and 2012. The heterogeneity of the tumor characteristics, for a given cancer, presents a great challenge in the research of new effective treatments. In this context, much hope is placed in the research of predictive (or treatment selection) biomarkers that reflect the patients' characteristics in order to guide treatment choice. For example, in the metastatic colorectal cancer setting, it is admitted that the addition of cetuximab (an anti-EGFR) to classical chemotherapy (the FOLFOX4), only improve the outcome of patients with *KRAS* wild-type tumors. In that context, the *KRAS* gene is a binary treatment selection marker, but plenty of biomarkers result from some quantifications or dosage measurements.

The first aim of this thesis is to quantify the global treatment selection ability of a biomarker. After a review of the existing litterature, a method based on an extension of ROC curves is proposed and compared to existing methods. Its main advantage is that it is non-parametric, and that it does not depend on the mean risk of event in each treatment arm.

In a second time, when a quantitative treatment selection biomarker is assessed, there is a need to estimate a marker thereshold value above which one treatment is preferred, and below which the other treatment is recommended. An approach that relies on the definition of a utility function is proposed in order to take into account both efficacy and toxicity of treatments when estimating the optimal threshold. A Bayesian method for the estimation of the optimal threshold is proposed.

Keywords: biomarker evaluation ; predictive biomarker ; treatment selection ; receiver operating characteristic curves ; optimal threshold ; utility function ; colorectal cancer

---

Intitulé et adresse du laboratoire

Laboratoire de Biométrie et Biologie Evolutive - Equipe Biostatistiques-Santé
162 avenue Lacassagne
69424 Lyon Cedex 03

Remerciements

Mes remerciements vont tout d'abord à mon directeur de thèse **Fabien Subtil** et à ma co-directrice de thèse **Muriel Rabilloud** qui m'ont accueilli au sein du service de Biostatistique des Hospices Civils de Lyon et qui m'ont encadré au cours de ces trois dernières années. Sans vos conseils, votre disponibilité et votre motivation je n'aurais probablement pas été capable de mener à bien ce projet. Merci pour vos très nombreuses relectures que ce soit d'articles, ou du présent manuscrit. Et surtout merci de m'avoir permis de choisir ce sujet qui m'a parfois agacé, mais ô combien passionné.

Je remercie également les rapporteurs de cette thèse, **Christine Binquet** et **Xavier Paoletti**, pour avoir accepté ce rôle ainsi que **Hélène Jacqmin-Gadda** et **Delphine Maucort-Boulch** pour avoir accepté de faire partie de ce jury.

Mes pensées vont également au **personnel de la Fédération Francophone de Cancérologie Digestive**, et notamment à **Karine Le Malicot** également membre de mon comité de pilotage, qui m'ont permis de travailler sur les données de l'essai clinique PETACC-8 et d'enrichir mon manuscrit d'illustrations concrètes.

Je remercie également les autres membres de mon comité de pilotage **Vivian Viallon**, **François Gueyffier** ainsi que **Vincent Balter** pour leurs remarques toujours pertinentes et leurs encouragements.

Mes remerciements vont bien sûr à la **Fondation pour la Recherche Médicale** dont le soutien financier m'a permis de travailler dans de bonnes conditions.

Je tiens aussi à remercier les **enseignants** qui m'ont formé tout au long de mon parcours. De ma professeur de mathématiques au lycée l'Oiselet de Bourgoin-Jallieu, **Annie André**, qui a su me redonner le goût des chiffres, en passant par les enseignants du DUT STID de l'IUT Lumière Lyon 2, je pense notamment à **Hélène Chanvillard**, et de la Licence professionnelle Santé (spécialité Biostatistique) de l'Université Grenoble Alpes, et particulièrement **Frédérique Leblanc** et **Frédérique Letué**, jusqu'aux enseignants, et pour la plupart collègues, du Master B3S de l'Université Claude Bernard Lyon 1, je vous remercie tous de m'avoir permis d'atteindre cet objectif. J'adresse des remerciements tout particuliers à **René Ecochard**, nos discussions en fin de Master 2, et vos conseils ont été plus que précieux pour moi.

Merci à la super équipe du secrétariat du service. Vous avez été d'une aide précieuse face à la complexité administrative de certaines procédures, et avez été extrêmement attentives à ce que tout se déroule pour le mieux pendant mon passage. Mille mercis **Michèle**, **Stéphanie** et **Clémentine**.

Merci **Amna** de m'avoir accueilli dans ton bureau pour former le meilleur bureau du service ! Je n'aurais pas pu espérer mieux. Evidemment un immense merci à tous les membres du service de Biostatistique pour leurs conseils et leur écoute, particulièrement pour les séances « salon de thé » improvisées dans notre bureau. Merci pour votre soutien et votre bonne humeur. Merci **Carole**, **Sylvain**, **Jean**, **Catherine**, **Fatima**, **Pascal** et **tous les autres**.

Parce que la vie de thésard n'est pas toujours facile à gérer, surtout lorsque j'étais confronté à mes doutes, il est parfois bon de pouvoir en parler avec des personnes qui vivent la même chose au quotidien. Merci aux deux autres thésards du service, **Mathieu** et **Joris**, d'avoir su apporter un regard différent sur mes problématiques, et surtout d'avoir été une vraie source de bonne humeur pour moi lors de mes passages à Lyon Sud.

D'après Calogero, « on n'est riche que de ses amis », alors c'est dit, et je sais que vous serez encore là après cette thèse. Merci aux membres de la tribu sopalin : **William**, **Florian**, **Stéphane** et **Jordan**. Merci infiniment à mon « couple à trois », **Mathilde** et **Elodie**, pour nos soirées mémorables. Merci pour votre soutien constant.

Enfin, et surtout, merci à **ma famille**, **mes parents** et **mon frère** (t'inquiète le sang je t'ai pas oublié), ainsi que ma compagne **Vanessa**. Sans votre soutien et votre affection, je ne serais probablement pas là où j'en suis aujourd'hui.

Cette longue aventure se termine ici. Ces dernières phrases me permettent de clore un chapitre de ma vie qui aura duré trois ans. Trois ans à travailler, comprendre, rechercher, mais également à m'amuser et à voyager des Pays-Bas à l'Espagne, en passant par l'Australie, ce qui m'aura permis de rencontrer des personnes enrichissantes autant sur le plan personnel que professionnel. Avant de conclure, je tiens à remercier **ceux que j'ai oubliés** mais qui n'en sont pas moins importants.

Cette thèse débute maintenant pour les **lecteurs les plus courageux** que je tiens eux aussi, et puisqu'il s'agit du maître mot, à remercier.

Production scientifique

Articles en lien avec la thèse

- Article publié :
Blangero Y, Rabilloud M, Ecochard R et Subtil F. *A Bayesian method to estimate the optimal threshold of a marker used to select patients' treatment*. Statistical Methods in Medical Research, 2019 ; doi : 10.1177/0962280218821394.

- Article en révision pour la revue Biometrical Journal :
Blangero Y, Rabilloud M, Laurent-Puig P, Le Malicot K, Lepage C, Ecochard R, Taïeb J et Subtil F. *The area between ROC curves, a non-parametric method to evaluate a biomarker for patient treatment selection*.

Communications orales

- EpiClin 2018, Nice, France
Blangero Y, Rabilloud M et Subtil F. L'aire entre les courbes, une méthode non-paramétrique pour évaluer la capacité prédictive globale d'un biomarqueur.

- ISCB 2018, Melbourne, Australie
Blangero Y, Rabilloud M et Subtil F. Optimal threshold estimation method for treatment selection markers.

- Séminaire LJK-Probabilités & Statistiques du 31 Janvier 2019, Grenoble, France
Blangero Y. Méthode bayésienne pour estimer le seuil optimal d'un marqueur utilisé pour choisir le traitement des patients.

Communications affichées

- ISCB 2017, Vigo, Espagne
A non-parametric method to evaluate predictive biomarkers for treatment selection : The area between curves.

- MEMTAB 2018, Utrecht, Pays-Bas

A non-parametric method to evaluate predictive biomarkers for treatment selection : The area between curves.

Notations

Général

Un vecteur est par convention un vecteur colonne.

n est la taille de l'échantillon = nombre de sujets.

$Y, X, Z, \ldots$: variables aléatoires à valeurs dans $\boldsymbol{R}$.

Les symboles désignant des vecteurs ou des matrices sont en caractère gras.

$\boldsymbol{\beta}'$: transposée du vecteur $\boldsymbol{\beta}$.

$y, x, z, \ldots$: valeurs réelles que peuvent prendre les variables aléatoires.

$F_X(x)$: fonction de répartition de la variable X (ou bien $F(x)$ s'il n'y a aucune ambiguïté).

$\mathcal{N}(\mu, \sigma^2)$: distribution gaussienne d'espérance μ et de variance σ^2.

Fonction indicatrice : $\mathbb{1}(Y = X) = 1$ si $Y = X$ et 0 sinon.

$\mathbf{E}(X)$: espérance de la variable aléatoire X.

Modèles de régression

Y : variable à expliquer ; X, Z variables explicatives.

$\theta, \beta, \gamma, \ldots$ et autres lettres grecques minuscules : paramètres.

log : logarithme népérien.

$\widehat{\theta}$: estimateur de θ. La vraie valeur du paramètre est désignée par θ.

Modèles de survie

T : temps jusqu'à la survenue de l'évènement d'intérêt.

C : temps jusqu'à la censure.

$Y = \min(T, C)$: temps de suivi observé.

$\delta = \mathbb{1}(T \leq C)$: indicatrice d'évènement.

$S(t) = 1 - F(t)$: fonction de survie .

$S_Y(t) = \mathrm{P}(Y > t)$, $S_T(t) = \mathrm{P}(T > t)$, $S_C(t) = \mathrm{P}(C > t)$: sont les différentes notations pour les fonctions de survie utilisées dans les développements.

Table des matières

Table des figures

Liste des tableaux

Introduction

La thématique générale de ce travail de recherche concerne la méthodologie de l'évaluation des biomarqueurs prédictifs quantitatifs et de leur utilisation en pratique clinique, dans le but d'adapter les thérapies disponibles pour une pathologie donnée aux caractéristiques des patients reflétées par ce type de biomarqueurs. La question est de savoir comment quantifier l'intérêt global de ces biomarqueurs pour guider le choix entre deux traitements, mais également de pouvoir définir une valeur seuil optimale de marqueur permettant de définir une règle de décision médicale. Par exemple, les patients ayant une valeur de marqueur supérieure au seuil optimal recevront plutôt une thérapie, tandis que les patients ayant une valeur de marqueur inférieure à ce seuil recevront un traitement différent.

Les méthodes de quantification de la capacité globale du marqueur à guider le choix du traitement peuvent être classées en deux familles distinctes : les méthodes s'appuyant sur le différentiel de réponse aux deux traitements à l'échelle individuelle pour évaluer de manière *directe* le différentiel moyen, et les méthodes n'ayant pas recours à cette mesure à l'échelle individuelle mais qui permettent de mesurer le différentiel moyen de manière *indirecte*. Ces deux familles sont soumises à des contraintes méthodologiques différentes et conduisent à des indicateurs dont l'interprétation est sensiblement différente, bien que l'objectif soit le même.

Les méthodes permettant d'estimer le seuil optimal du marqueur sont beaucoup moins développées dans la littérature. L'estimation de ce seuil ne se fait pas seulement en tenant compte de l'efficacité de chacun des traitements, mais également en évaluant le bénéfice et le coût de chacun d'entre eux en matière d'état de santé.

Les méthodes développées dans le cadre de cette thèse sont appliquées à des données issues de l'essai clinique PETACC-8 qui comparait l'efficacité d'une chimiothérapie classique à une combinaison de chimiothérapie et d'anti-EGFR (Epidermal Growth Factor Receptor) pour le traitement du cancer colorectal de stade III réséqué. L'étude n'a pas pu mettre en évidence de différence d'efficacité significative entre les deux options de traitement, les investigateurs ont donc cherché à identifier des sous-groupes de patients retirant un bénéfice de l'ajout d'anti-EGFR. L'objectif a donc été d'identifier des marqueurs ayant une capacité à guider le choix du traitement suffisamment importante pour avoir un intérêt clinique, et également de définir un seuil optimal permettant leur utilisation en pratique clinique.

Le premier chapitre de cette thèse décrit les problématiques ayant motivé ce travail de recherche et pose le contexte clinique des développements méthodologiques qui suivent.

Le deuxième chapitre effectue une revue de la littérature des méthodes existantes pour quanti-

fier la capacité prédictive globale d'un biomarqueur, et présente un nouvel indicateur au travers d'un article soumis au Biometrical Journal, ainsi que des compléments à l'article.

Enfin le troisième chapitre présente une méthode d'estimation du seuil optimal d'un marqueur prédictif décrite dans un article publié dans la revue Statistical Methods in Medical Research, ainsi que des compléments à l'article et des perspectives de développement.

Chapitre 1

Contexte

1.1 Zoom sur la cancérologie

De nos jours, la présence très marquée des cancers impose de nombreux défis en matière de santé publique. En France, le nombre de nouveaux cas de cancers a augmenté de 77 % entre 1990 et 2018. Cette augmentation est liée en partie à l'accroissement de la population, à son vieillissement, ainsi qu'à une hausse du risque de développer un cancer (Defossez et al., 2019). Le cancer de la prostate est le plus fréquent chez les hommes tandis que chez les femmes il s'agit du cancer du sein, et ce quel que soit l'âge. Enfin, chez les hommes et les femmes, le cancer colorectal représente environ 10 % des cas de cancer.

Le dépistage du cancer et le développement de stratégies thérapeutiques permettant de lutter face à cette maladie représentent donc de véritables défis sociétaux et des enjeux de santé publique.

L'évaluation de l'efficacité des traitements développés en oncologie nécessite l'analyse statistique de données, recueillies lors d'essais cliniques, en mesurant pour chaque patient un critère de jugement principal qui reflète cette efficacité, comme par exemple le temps jusqu'à la récidive du cancer, ou bien jusqu'au décès.

L'analyse de ces données permet de comparer de nouvelles stratégies thérapeutiques, dites « innovantes », à des stratégies de « référence » qui ont déjà fait leur preuve dans la pratique, au travers de tests statistiques.

Dans le cadre de cette thèse, une attention particulière sera apportée au cas du cancer colorectal, et notamment à l'analyse des données de l'essai PETACC-8 qui évalue le risque de récidive chez des patients de stade III après résection de leur tumeur (Taieb et al., 2014).

1.1.1 Cancérogénèse

Le développement du cancer au sein de l'organisme ne se fait pas de manière instantanée, il s'agit d'un processus long et d'une combinaison d'évènements conduisant à la transformation d'un tissu physiologique en un tissu cancéreux. Il est possible de découper ce processus en plusieurs phases présentées ci-après.

Dans notre vie quotidienne nous (et nos cellules) sommes exposés à des agents carcinogènes qu'ils soient physiques (rayonnement ionisant, ultraviolet, ...), chimiques (tabac, alcool, ...) ou bien viraux (virus d'Epstein-Barr, papillomavirus, ...). Cette exposition va entraîner une altération ponctuelle définitive du patrimoine génétique de la cellule exposée (mutation, délétion, translocation), on parle alors de mutation *initiatrice*. Cette modification ponctuelle n'est pas suffisante à elle seule pour transformer la cellule exposée en cellule cancéreuse, cependant elle va donner à la cellule un avantage en matière de prolifération ou de survie car elle concerne souvent des gènes majeurs responsables de la réparation de l'ADN ou de l'apoptose (mécanisme d'autodestruction de la cellule). On notera que pour les cancers d'origine génétique, cette mutation initiatrice est déjà présente à la naissance. A ce stade, il n'y a aucun symptôme visible et aucune modification phénotypique.

Après cette phase d'initiation, va suivre la phase de *promotion*. Cette phase est liée à l'exposition prolongée de la cellule à des agents promoteurs (facteurs de croissance, cytokines, hormones), c'est-à-dire des agents qui vont favoriser la multiplication cellulaire, et ainsi l'expression de la mutation initiatrice. Lors de ces multiplications, les mutations génétiques vont s'accumuler à chaque division jusqu'à former la cellule cancéreuse.

La phase de *progression* qui suit la phase de promotion consiste en la prolifération des cellules anormales favorisée par la perte du contrôle du cycle cellulaire (à cause de la mutation des gènes régissant la réparation de l'ADN d'une cellule, ou bien régissant le mécanisme de l'apoptose par exemple) et aboutissant à la formation d'un clone tumoral (il s'agit d'une accumulation de cellules ayant les mêmes anomalies génétiques). Cette phase aboutit rapidement à la formation d'une lésion précancéreuse constituée de plusieurs clones tumoraux expliquant alors l'hétérogénéité génétique de la lésion. Cette lésion précancéreuse est strictement intra-épithéliale, ce qui signifie qu'à ce stade elle est compartimentée.

La poursuite de la prolifération tumorale aboutit au développement localisé d'une tumeur infiltrante, c'est la phase d'*invasion*. On assiste alors à un envahissement de l'organe concerné ainsi que des organes voisins. Enfin, les cellules cancéreuses disséminent dans tout l'organisme par différentes voies selon la localisation initiale (voies lymphatiques, voies hématogènes), cette phase de *dissémination* aboutit alors à la formation de métastases dans des organes éloignés de la tumeur initiale.

1.1.2 Evolution du cancer colorectal en France

Le cancer colorectal est un cancer qui se développe dans le gros intestin. La Figure 1.1 permet de visualiser l'emplacement du côlon qui représente la plus grosse partie du gros intestin, ainsi que le rectum qui est la dernière partie du gros intestin. Généralement, les premiers symptômes de ce cancer apparaissent à un stade plutôt avancé de la maladie. Les symptômes les plus courants sont des modifications du transit intestinal, des douleurs abdominales, la fatigue et la perte de poids. La présence de sang dans les selles est également un indicateur important pour le dépistage de la maladie. Néanmoins, le diagnostic final affirmant le cancer n'est rendu qu'après une analyse anatomopathologique de la tumeur.

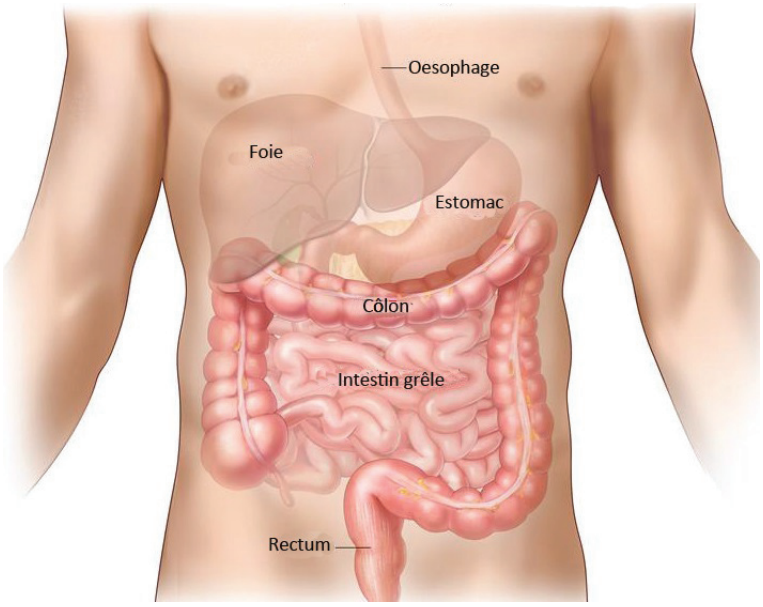


FIGURE 1.1 – Anatomie du système digestif

En France, on estimait à 43 336 le nombre de nouveaux cas de cancers colorectaux en 2018, dont 54 % survenant chez l'homme. Le sex-ratio hommes/femmes du taux d'incidence standardisé monde est estimé à 1.4. Les taux de mortalité standardisés monde estimés en 2018 étaient de 11.5 chez l'homme et 6.9 chez la femme (Defossez et al., 2019). La Figure 1.2 présente la tendance chronologique des taux d'incidence et de mortalité standardisés monde chez les hommes et les femmes entre 1990 et 2018 (Defossez et al., 2019).

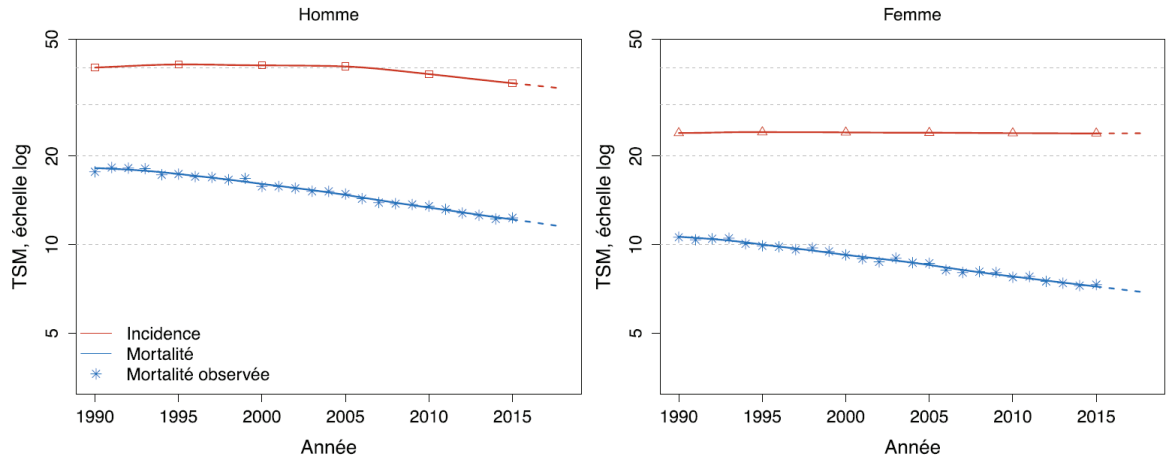


FIGURE 1.2 – Evolution du cancer colorectal en France
TSM : Taux standardisés monde (source : Defossez et al. (2019))

On observe depuis 2005 une baisse de l'incidence chez les hommes accompagnée d'une diminution du taux de mortalité. Chez les femmes, le taux d'incidence reste constant depuis 1990, et est accompagné d'une diminution du taux de mortalité. L'évolution différente de l'incidence et de la mortalité s'explique par le développement de méthodes de dépistage plus performantes ainsi que par l'amélioration de la prise en charge thérapeutique des patients. Depuis 2005, on

observe une diminution de l'incidence qui peut également être associée à un accès au dépistage plus facile permettant la résection des lésions précancéreuses de manière plus précoce.

La prise en charge thérapeutique des patients diffère selon le stade du cancer. Au moment du diagnostic, on classe le cancer selon le système de classification TNM proposé par l'UICC (Union for International Cancer Control), afin de déterminer quel est le stade du cancer.

La catégorie « T » permet de décrire le site initial de la tumeur, la catégorie « N » décrit l'implication des ganglions lymphatiques régionaux dans le développement de la tumeur, et la catégorie « M » décrit la présence ou non de métastases à distance du site primitif de la tumeur. Les Tableaux 1.1, 1.2 et 1.3 présentent les catégories TNM pour le cancer colorectal.

Tableau 1.1 – Classification T du cancer colorectal

Classification	Description
TX	Renseignements insuffisants pour classer la tumeur primitive
T0	Pas de signe de tumeur primitive
Tis	Carcinome in situ : intra-épithélial ou envahissant la lamina propria (chorion de la muqueuse)
T1	Tumeur envahissant la sous-muqueuse
T2	Tumeur envahissant la musculeuse
T3	Tumeur envahissant la sous-séreuse ou les tissus péri-coliques ou péri-rectaux non péritonisés
T4	Tumeur envahissant directement d'autres organes ou d'autres structures et/ou perforant le péritoine viscéral **T4a** : tumeur perforant le péritoine viscéral **T4b** : tumeur envahissant directement d'autres organes ou d'autres structures

Tableau 1.2 – Classification N du cancer colorectal

Classification	Description
NX	Renseignements insuffisants pour classer les ganglions lymphatiques régionaux
N1	Métastases dans un à trois ganglions lymphatiques régionaux **N1a** : Métastases dans un seul ganglion régional **N1b** : Métastases dans deux à trois ganglions lymphatiques régionaux **N1c** : Nodules tumoraux dans la sous-séreuse ou dans les tissus mous non péritonéalisés péri-coliques ou péri-rectaux sans atteinte ganglionnaire lymphatique
N2	Métastases dans plus de trois ganglions lymphatiques régionaux **N2a** : Métastases dans quatre à six ganglions lymphatiques régionaux **N2b** : Métastases dans sept ou plus ganglions lymphatiques régionaux

Les combinaisons des catégories TNM permettant de déterminer le stade du cancer sont présentées dans le Tableau 1.4.

Tableau 1.3 – Classification M du cancer colorectal

Classification	Description
M0	Pas de métastase à distance
M1	Présence de métastases à distance **M1a** : Métastases dans un seul organe (foie, poumon, ovaire, ganglions lymphatiques non régionaux) **M1b** : Métastases dans plus d'un organe ou dans le péritoine

Tableau 1.4 – Détermination du stade du cancer colorectal

Stade	T	N	M
Stade 0	Tis	N0	M0
Stade I	T1, T2	N0	M0
Stade II	T3, T4	N0	M0
Stade IIA	T3	N0	M0
Stade IIB	T4a	N0	M0
Stade IIC	T4b	N0	M0
Stade III	Quelque soit T	N1, N2	M0
Stade IIIA	T1, T2	N1	M0
	T1	N2a	M0
Stade IIIB	T3, T4a	N1	M0
	T2, T3	N2a	M0
	T1, T2	N2b	M0
Stade IIIC	T4a	N2a	M0
	T3, T4a	N2b	M0
	T4b	N1, N2	M0
Stade IVA	Quelque soit T	Quelque soit N	M1a
Stade IVB	Quelque soit T	Quelque soit N	M1b

Pour chacun des stades du cancer, la prise en charge du patient ne sera pas la même, c'est d'ailleurs tout l'intérêt de cette classification. Il est possible de présenter de manière succincte et simplifiée les stratégies thérapeutiques selon les stades du cancer de la manière suivante :

- Au stade 0, la tumeur ne touche que des zones superficielles de l'intestin. La tumeur est alors retirée au cours d'une chirurgie.
- Au stade I, le cancer touche une zone plus profonde des muqueuses, et il se peut même que les muscles du côlon et du rectum soient impactés. La tumeur doit alors être retirée chirurgicalement ainsi que les ganglions lymphatiques locaux.
- Au stade II, les muscles de l'intestin, ainsi que les organes proches, sont touchés. Le traitement consiste alors en une ablation chirurgicale de tous les tissus touchés ainsi que d'une chimiothérapie chez les patients atteints d'un cancer du côlon. Une radiothérapie, ou bien une radiothérapie combinée à une chimiothérapie est nécessaire pour les patients atteints d'un cancer du rectum.
- Au stade III, les organes adjacents au côlon sont touchés, ainsi que les ganglions lymphatiques régionaux. Le traitement est le même que pour le stade II.
- Au stade IV, des organes éloignés sont également touchés (généralement le foie et les poumons), le cancer est en phase de généralisation. Les traitements consistent alors en des combinaisons de chimiothérapies et de thérapies biologiques ciblées dans le but de réduire la taille des métastases afin de les rendre, si possible, opérables.

La chirurgie demeure le principal acte curatif du cancer colorectal. Cependant, depuis les années 1990, l'utilisation de chimiothérapie adjuvante basée sur le fluorouracil, chez les patients ayant un cancer colorectal de stade III après résection de la tumeur par chirurgie, a démontré son effet dans la diminution du risque de récidive (Mamounas et al., 1999; Gill et al., 2004). Un essai plus récent a démontré l'intérêt de l'ajout d'oxaliplatine aux chimiothérapies basées sur le fluorouracil (André et al., 2009) dans la diminution du risque de rechute, et ces résultats ont été confirmés lors d'un autre essai (Kuebler et al., 2007; Yothers et al., 2011). On constate donc un véritable effort de la recherche ces dernières années pour améliorer la prise en charge et le devenir de l'ensemble des patients atteints d'un cancer colorectal.

1.2 L'ère des biomarqueurs

Le terme « biomarqueur », issu de la combinaison des termes « marqueur » et « biologique », est désormais ancré dans le langage courant en pratique médicale ainsi qu'en recherche clinique. La définition actuellement admise par la communauté scientifique caractérise les biomarqueurs comme « des caractéristiques mesurables objectivement et évaluées comme des indicateurs de processus biologiques normaux, pathogènes, ou bien des réponses pharmacologiques à une intervention thérapeutique » (Biomarker Definitions Working Group, 2001). Les biomarqueurs les plus connus, et les plus accessibles, sont des caractéristiques biologiques qui peuvent être détectées et mesurées dans certaines parties de l'organisme notamment dans des tissus spécifiques

ou bien des fluides biologiques (ex : la peau, le sang, l'urine, ...). Ainsi, l'expression de certains gènes, les produits issus d'anticorps, ou bien certaines hormones peuvent être des exemples de biomarqueurs.

La mesure et le suivi de biomarqueurs sont désormais réalisés dans de nombreuses spécialités, et tout particulièrement en oncologie (Schiffer, 2009), dans le domaine des maladies cardiovasculaires (Gerszten and Wang, 2008), ainsi que des maladies infectieuses (Tajik et al., 2013). Des exemples classiques de biomarqueurs sont le dosage des antigènes CA19-9 ou CA125 pour le diagnostic de certains cancers, ou bien la suractivité de la tyrosine kinase BCR-ABL chez les patients atteints de leucémie myéloïde chronique qui a permis le développement de l'une des premières thérapies ciblées (Imatinib) dans le but d'inhiber cette sur-activité (Cohen et al., 2002; O'Brien et al., 2003; Hochhaus et al., 2017).

Les biomarqueurs peuvent donc être utilisés pour des objectifs très divers, et il est possible d'établir une classification des biomarqueurs en fonction de leur domaine d'application en pratique clinique présentée dans le Tableau 1.5 (Carlomagno et al., 2017).

Tableau 1.5 – Définition des différents types de biomarqueur

Type de biomarqueur	Définition
Thérapeutique	Généralement une protéine utilisée comme cible d'une thérapie
Diagnostique	Marqueur biologique permettant de détecter la présence d'une maladie donnée
Pronostique	Marqueur biologique permettant de prédire l'évolution de la maladie dans le temps, ainsi que le devenir des patients
Prédictif	Marqueur biologique permettant de prédire un différentiel de réponses entre deux thérapies afin de définir des sous-populations de patients à même de retirer un bénéfice plus important d'une thérapie spécifique

Les biomarqueurs *thérapeutiques* sont généralement des protéines utilisées comme cible d'une thérapie, ils sont donc le reflet des mécanismes d'action de la maladie et sont de fait cruciaux dans le développement de nouvelles stratégies thérapeutiques qui peuvent alors cibler ces biomarqueurs pour empêcher le développement de la maladie, ou bien pour l'éradiquer.

Les biomarqueurs *diagnostiques* permettent de détecter la présence d'une maladie donnée chez un patient. Généralement ces biomarqueurs permettent de confirmer un diagnostic lorsque sa mesure est mise au regard des symptômes observés chez le patient. Ils sont particulièrement utiles pour éviter à certains patients de subir des examens invasifs (par exemple une biopsie).

Les biomarqueurs *pronostiques* sont utiles lorsque le diagnostic de la maladie a déjà été posé et que l'on souhaite avoir des informations concernant l'évolution future de cette maladie dans le temps. C'est ce type de biomarqueurs qui est généralement utilisé pour prédire le risque de récidive d'un cancer.

Enfin les biomarqueurs *prédictifs* sont utiles pour choisir entre deux thérapies pour une maladie donnée. Ils permettent de prédire la réponse d'un patient à chacune des thérapies envisagées et donc de donner à ce patient le traitement qui lui sera le plus efficace. ou le plus utile clinique-

ment. L'émergence de ces biomarqueurs dans la pratique clinique s'inscrit donc logiquement dans le développement de la médecine de précision (ou bien médecine personnalisée). Par la suite, lorsque l'on dira qu'un patient retire un bénéfice plus important du traitement innovant, par rapport au traitement de référence, cela signifiera que sa réponse au traitement innovant est meilleure que celle au traitement de référence.

Parfois, la simple présence (ou absence) d'un marqueur dans l'organisme peut être révélatrice de la maladie, ou bien de son pronostic. On parle dans ces cas-là de marqueurs *binaires*. Dans le cas d'un biomarqueur prédictif binaire, les patients pour lesquels le biomarqueur est présent auront un bénéfice à recevoir l'un des deux traitements envisagés, alors que ceux pour qui le biomarqueur est absent retireront un bénéfice de la deuxième option thérapeutique. Dans d'autres cas, c'est une concentration plus ou moins élevée du biomarqueur dans l'organisme qui donne une indication sur la maladie, son pronostic, ou bien sur la capacité du biomarqueur à guider le choix du traitement. On parle alors de biomarqueur *quantitatif*.

Pour une valeur donnée du biomarqueur quantitatif, il est possible d'attribuer un risque d'avoir la maladie (biomarqueur diagnostique), un risque pour la maladie de changer d'état (biomarqueur pronostique), ou bien une différence de risques d'évolution de la maladie entre deux thérapies (biomarqueur prédictif).

L'analyse des biomarqueurs quantitatifs nécessite la définition d'un seuil au-dessus (ou en-dessous) duquel la concentration du biomarqueur est jugée anormale (biomarqueur diagnostique), le risque pour la maladie de changer d'état suffisamment élevé (biomarqueur pronostique), ou bien que la différence de risques observée pour un patient donné est suffisamment grande pour justifier l'administration d'une thérapie plutôt qu'une autre (biomarqueur prédictif).

Comme expliqué précédemment, l'émergence des biomarqueurs prédictifs dans la pratique médicale est fortement liée au développement de la médecine de précision. Celle-ci s'appuie sur l'analyse des caractéristiques génétiques ou biologiques des patients ainsi que de leur tumeur afin de proposer des stratégies thérapeutiques personnalisées et adaptées à chaque patient. L'objectif n'est donc plus de développer de nouvelles thérapies qui seront efficaces pour tous, mais des thérapies dont l'efficacité pourra être démontrée chez certaines catégories de patients. La médecine de précision trouve particulièrement sa place en cancérologie avec l'accès grandissant à l'analyse génétique des tumeurs afin de détecter quelles sont les mutations de gènes associées à l'évolution de la tumeur selon le site. La détection de ces mutations génétiques peut alors permettre le développement de thérapies ciblées qui vont tenter d'inhiber la sur-expression de certains gènes et donc de freiner l'évolution des tumeurs.

Pour reprendre l'exemple du cancer colorectal, plusieurs études ont démontré l'intérêt de l'ajout d'un inhibiteur de l'EGFR (Epidermal Growth Factor Receptor) aux combinaisons de chimiothérapies classiques pour des cancers de stade IV (Douillard et al., 2010; Bokemeyer et al., 2011; Douillard et al., 2013). Cependant, les patients présentant des tumeurs avec une mutation de l'exon 2 du gène *KRAS* ne retiraient aucun bénéfice de cette combinaison de traitement. D'autres études ont également démontré que, chez des patients ne répondant pas à la

chimiothérapie classique, les tumeurs ayant une mutation de l'exon 2 du gène *KRAS* étaient résistantes aux anticorps EGFR (Di Fiore et al., 2007; Lièvre et al., 2008; De Roock et al., 2008). Ainsi, l'identification de cette mutation génétique permet de choisir une stratégie thérapeutique adaptée aux caractéristiques de la tumeur du patient. Il s'agit d'un biomarqueur prédictif binaire car il permet de guider le choix du traitement.

Dans un monde idéal, on souhaiterait pouvoir évaluer le bénéfice thérapeutique à l'échelle individuelle, c'est-à-dire qu'on aimerait savoir, pour une personne donnée, quel est le traitement le plus efficace. Il est possible d'évaluer le bénéfice individuel dans des essais dits en « cross-over », où chaque patient pourra recevoir chacun des traitements évalués. Cependant, en cancérologie, ce type d'essai est plus difficile à mettre en place en raison du caractère évolutif de la maladie qui va rendre la comparaison d'efficacité des différents traitements non interprétable. Ce besoin d'évaluer le bénéfice d'une thérapie à l'échelle individuelle constituera la base du développement de plusieurs méthodes qui seront détaillées par la suite.

Chaque jour des publications scientifiques annoncent la découverte de nouveaux biomarqueurs et apportent ainsi leur lot d'espoirs pour la prise en charge de nombreuses maladies. Cependant, il est nécessaire de prendre un recul suffisant sur ces découvertes car il s'avère que bon nombre de ces biomarqueurs se révèlent peu utiles en pratique clinique lorsqu'ils sont utilisés à l'échelle de la population générale (Ransohoff, 2004; Wong and Pollock, 2014).

Pour la suite de ce travail, le terme « marqueur » sera principalement utilisé, puisqu'un marqueur peut être un biomarqueur, une combinaison de biomarqueurs ou bien des combinaisons de caractéristiques des patients ne correspondant pas à la définition d'un biomarqueur.

1.3 L'essai clinique PETACC-8

L'essai clinique PETACC-8 est un essai de phase III international, contrôlé, et randomisé qui compare l'efficacité d'une chimiothérapie à base de FOLFOX4 à une combinaison de FOLFOX4 + cetuximab (un inhibiteur de l'EGFR) chez des patients âgés de 18 à 75 ans ayant un cancer colorectal de stade III confirmé et réséqué au début de l'étude. Comme les patients ayant une mutation de l'exon 2 du gène *KRAS* sont résistants aux anti-EGFR dans le contexte du cancer colorectal métastatique, un amendement au protocole a restreint l'inclusion à des patients avec des tumeurs n'ayant pas cette mutation. Le critère de jugement étudié était la survie sans maladie après résection de la tumeur, tenant compte comme événement d'une récidive locorégionale ou métastatique, de l'apparition d'un second cancer du côlon ou du rectum, ou du décès (Taieb et al., 2014).

1.3.1 Sélection des patients

La Figure 1.3 présente le processus de sélection des patients dans l'essai PETACC-8. Ici le processus a été simplifié, le diagramme complet est disponible dans l'article de Taieb et al. (2014).

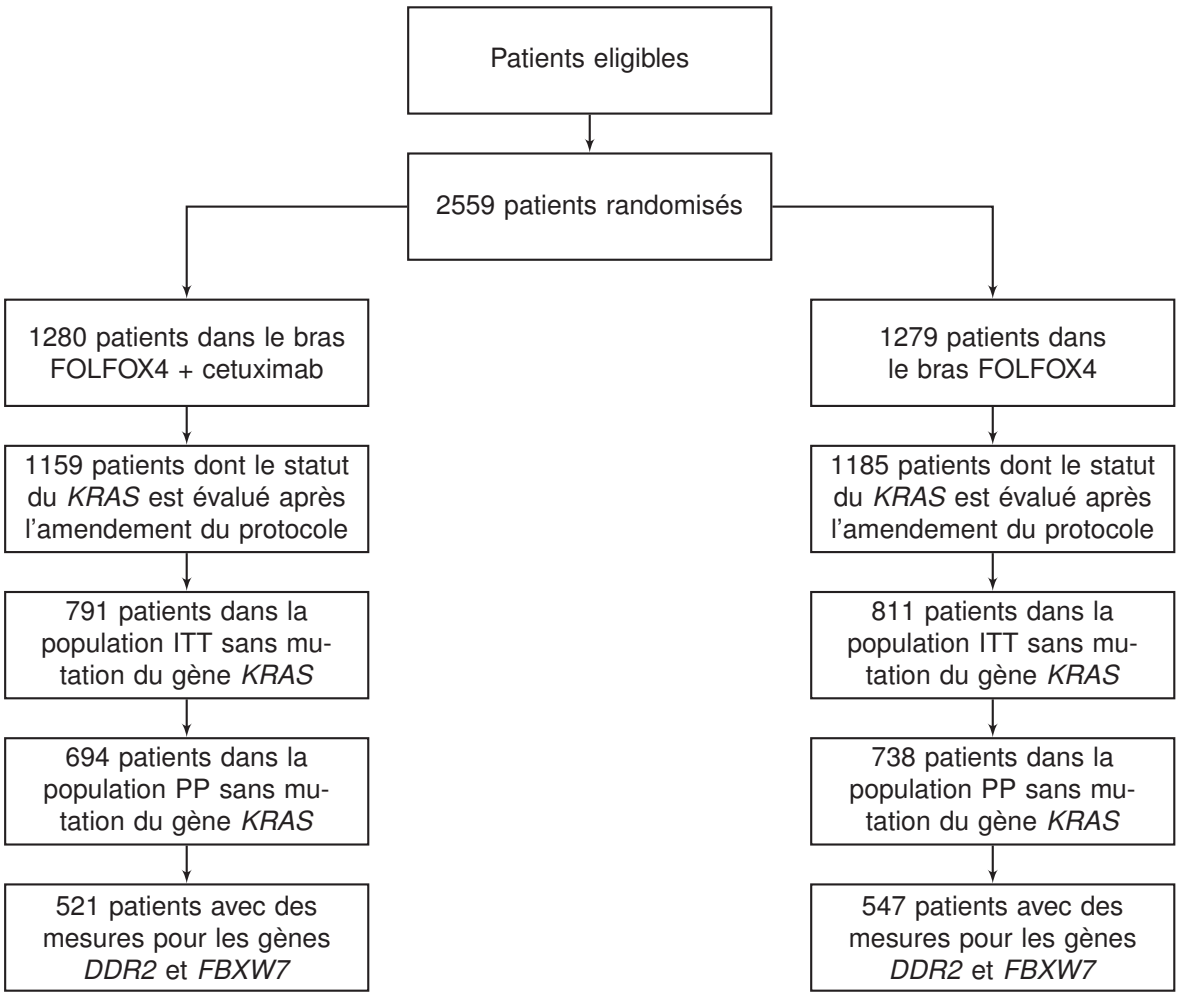


FIGURE 1.3 – Profil de l'essai PETACC-8

Au départ de l'essai, 2559 patients avaient été randomisés entre les bras de traitement FOLFOX4 ($n = 1279$) et FOLFOX4 + cetuximab ($n = 1280$). L'amendement au protocole a restreint la population en intention de traiter (ITT) aux 1602 patients sans mutation du gène *KRAS*, dont 811 étaient dans le bras FOLFOX4 et 791 dans le bras FOLFOX4 + cetuximab.

La population Per Protocol (PP) a été définie en excluant de la population ITT :

- les patients qui n'ont pas démarré leur traitement,
- les patients qui ont suivi leur traitement moins de 8 semaines,
- les patients pour lesquels des déviations majeures au protocole ont été observées.

Ces exclusions ont ainsi restreint la population PP à 1432 patients, dont 738 étaient dans le bras FOLFOX4 et 694 dans le bras FOLFOX4 + cetuximab.

1.3.2 Description de la population Per Protocol

Les caractéristiques de la population Per Protocol de l'essai sont présentées dans le Tableau 1.6. L'âge, la localisation et la classification de la tumeur, ainsi que la présence d'obstruction ou de perforation de l'intestin constituent des facteurs pronostiques importants dans le cadre du cancer colorectal.

Tableau 1.6 – Description de la population Per Protocol de l'essai PETACC-8

Caractéristiques	FOLFOX4 ($n = 738$)	FOLFOX4 + cetuximab ($n = 694$)
Sexe		
Homme	425 (57.6 %)	409 (58.9 %)
Femme	313 (42.4 %)	285 (41.1 %)
Age		
Médiane	60	60
Q1 - Q3	[53 - 66]	[52 - 66]
Etendue	[21 - 75]	[19 - 75]
Localisation tumorale		
Gauche	475 (64.4 %)	443 (63.8 %)
Droite	257 (34.8 %)	245 (35.3 %)
Les deux	3 (0.4 %)	5 (0.7 %)
Manquant	3 (0.4 %)	1 (0.1 %)
Classification T		
T1-T2-T3	611 (82.8 %)	551 (79.4 %)
T4	126 (17.1 %)	141 (20.3 %)
Manquant	1 (0.1 %)	2 (0.3 %)
Classification N		
N0-N1	472 (64.0 %)	436 (62.8 %)
N2	266 (36.0 %)	258 (37.2 %)
Obstruction et perforation de l'intestin		
Obstruction et/ou perforation	133 (18.0 %)	135 (19.5 %)
Aucune des deux	605 (82.0 %)	559 (80.5 %)

Dans le bras FOLFOX4, 64.4 % des patients avaient une tumeur localisée dans le côlon gauche (475 patients) et 34.8 % dans le côlon droit (257 patients), contre 63.8 % (443 patients) et 35.3 % (245 patients) dans le bras FOLFOX4 + cetuximab, respectivement. Concernant les classifications T et N, 82.8 % des patients dans le bras FOLFOX4 étaient classés en T1, T2 ou T3 (611 patients), contre 79.4 % dans le bras FOLFOX4 + cetuximab (551 patients), et 64 % des patients du bras FOLFOX4 étaient classés dans les catégories N0 ou N1 (472 patients), contre 62.8 % dans le bras FOLFOX4 + cetuximab (436 patients). Enfin, il est à noter que 18 % des patients présentaient une obstruction et/ou une perforation du côlon dans le bras FOLFOX4 (133 patients), contre 19.5 % dans le bras FOLFOX4 + cetuximab (135 patients). Globalement, l'étude n'a pas montré de différence de survie entre les deux traitements, que ce soit dans la population en ITT (HR = 1.047 [0.853 ; 1.286]) ou dans la population PP (HR = 1.019 [0.816 ; 1.274]) d'où l'intérêt de rechercher d'éventuels sous-groupes de patients qui pourraient retirer un bénéfice de l'ajout du cetuximab.

1.3.3 Niveaux d'amplification des gènes *DDR2* et *FBXW7*

L'amplification d'un gène est un phénomène désormais bien connu en oncologie. Il s'agit d'un processus impliquant l'amplification de la séquence d'ADN codant le gène en question, entraînant alors un dérèglement dans la production de certaines protéines par les cellules de l'organisme.

Dans le cadre de l'essai PETACC-8, de nombreux tests ont été menés pour identifier des gènes dont la séquence codante aurait pu être amplifiée du fait du cancer, et de voir si les traitements évalués dans le cadre de l'essai permettaient d'inhiber la sur-expression induite de ces gènes.

Au total, ce sont 22 niveaux d'amplification de gènes différents qui ont été étudiés. Deux de ces marqueurs génétiques ont été retenus pour illustrer les différentes méthodes statistiques détaillées par la suite. Le premier d'entre eux est le niveau d'amplification du gène *DDR2* qui est un marqueur prometteur dans l'identification d'une hétérogénéité dans la différence de réponse aux deux traitements étudiés dans le cadre de PETACC-8. Le second est le niveau d'amplification du gène *FBXW7* pour lequel des résultats bien moins bons sont attendus.

L'étude de ces gènes a donc de nouveau restreint la population Per Protocol à la population de patients pour lesquels les niveaux d'amplification de ces deux gènes ont pu être mesurés. Après cette dernière sélection, il restait donc au total 1068 patients pour l'évaluation des capacités prédictives de ces deux gènes, dont 547 étaient randomisés dans le bras FOLFOX4, et 521 dans le bras FOLFOX4 + cetuximab.

Un tel processus de sélection pourrait remettre en cause l'hypothèse de randomisation de l'essai clinique, nécessaire dans l'analyse des capacités prédictives de ces marqueurs. La Figure 1.4 représente les densités de probabilité des deux marqueurs génétiques par bras de traitement. On constate que le processus de sélection ne semble avoir que peu impacté l'hypothèse de randomisation car les densités de probabilité de chaque marqueur sont assez similaires par bras de traitement.

Il est à noter également que ce processus de sélection n'a que très peu impacté la courbe

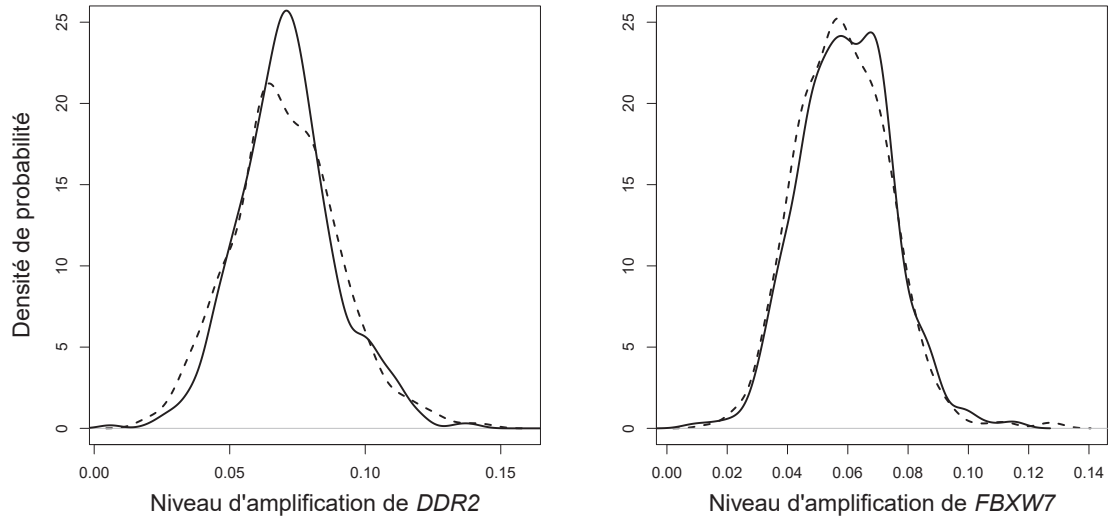


FIGURE 1.4 – Distributions des niveaux d'amplification des gènes *DDR2* et *FBXW7* par bras de traitement
Courbes pleines : FOLFOX4 + cetuximab ; Courbes en pointillés : FOLFOX4

de survie observée dans chaque bras de traitement et qu'il n'y avait toujours aucune différence statistiquement significative entre les deux courbes de survie.

Dans le cadre de cette thèse, à visée d'illustration, il a été choisi de quantifier le caractère prédictif global des niveaux d'amplification des gènes *DDR2* et *FBXW7*. Une fois que cette capacité a été quantifiée, il a fallu choisir un seuil optimal de marqueur au-delà duquel l'un des deux traitements est préféré, et en-dessous duquel le deuxième traitement est recommandé. Ces objectifs, qui peuvent paraître simples au premier abord, soulèvent de nombreuses questions méthodologiques qui seront abordées par la suite.

Au regard des nombreux tests réalisés, ainsi que du caractère même des marqueurs étudiés, les résultats obtenus pour illustrer les méthodes sont à évaluer avec recul. En effet, la mesure des niveaux d'amplification des gènes est complexe et parfois sujette à des erreurs qui pourraient influencer l'évaluation des marqueurs étudiés dans le cadre de l'essai PETACC-8 (Nedelman et al., 1992). Les résultats obtenus dans le cadre de cet essai devront donc être confirmés par d'autres études.

Chapitre 2

Evaluer la capacité prédictive globale d'un marqueur

Le développement et les espoirs importants placés dans la médecine de précision ont favorisé le développement de méthodes statistiques et mathématiques permettant de mesurer la capacité d'un marqueur à guider le choix du traitement (appelée ci-après « capacité prédictive »).

On distingue plusieurs étapes dans l'analyse des marqueurs prédictifs. La première d'entre elles est une phase exploratoire dont l'objectif est d'identifier des marqueurs utiles pour guider le choix du traitement. Dans le cas des marqueurs quantitatifs, une deuxième étape d'analyse est nécessaire pour estimer un seuil du marqueur au-delà duquel un traitement sera préféré, et en-dessous duquel le deuxième traitement sera préféré. Le choix de ce seuil sera fait en évaluant l'impact de cette règle de décision dans la population avec l'objectif de maximiser le nombre de patients qui recevront un traitement efficace pour eux. Enfin, une dernière étape peut être identifiée dont l'objectif est la validation de l'utilisation du marqueur prédictif en pratique clinique au travers d'essais cliniques dont le design est construit dans ce but (Sargent et al., 2005). Cette troisième étape ne sera pas abordée dans le cadre de cette thèse.

2.1 Définition mathématique d'un marqueur prédictif

Comme expliqué précédemment dans le Chapitre 1, l'objectif lorsque l'on recherche un marqueur prédictif est d'identifier des sous-groupes de patients pour lesquels la différence d'efficacité entre les deux traitements varie. On parle d'hétérogénéité dans l'effet des traitements.

2.1.1 Notations

On note Z, le traitement réellement assigné au patient dans une étude, prenant comme valeur 0 (traitement de référence) ou 1 (traitement innovant), Y le critère de jugement principal permettant d'évaluer l'efficacité des traitements, et X le marqueur dont on souhaite évaluer le caractère prédictif, mesuré avant l'administration du traitement (cela signifie que la valeur du marqueur n'est pas influencée par l'un des deux traitements évalués). On note également $Y^{(z)}$

le résultat potentiel du critère de jugement pour le patient avec le traitement $z \in \{0,1\}$. Il est alors possible d'exprimer la différence d'efficacité moyenne entre les deux traitements pour une valeur donnée du marqueur X, notée $\delta(X)$, comme :

$$\delta(X) = \mathbf{E}[Y^{(1)} - Y^{(0)}|X]\,. \tag{2.1}$$

Lorsqu'il n'existe pas d'hétérogénéité dans l'effet des traitements sachant un marqueur X étudié, cela revient à dire que la différence d'efficacité entre les deux traitements est indépendante de X, alors on peut écrire :

$$\delta(X) = \mathbf{E}[Y^{(1)} - Y^{(0)}] = \Delta\,, \tag{2.2}$$

ce qui signifie que pour un marqueur non prédictif, la différence d'efficacité entre les deux bras de traitements est constante en fonction de X et égale à Δ.

Ainsi, plus les variations dans la fonction $\delta(X)$ sont importantes, plus il est facile d'identifier des sous-groupes de patients pour lesquels la différence d'efficacité entre les deux traitements n'est pas la même. Cette faculté à identifier des sous-groupes de patients est ce qui caractérise un marqueur prédictif. Ainsi, plus les variations de $\delta(X)$ sont importantes en fonction de X, plus la capacité prédictive du marqueur est grande.

Pour la suite de ce travail, le cas des critères de jugement binaires sera particulièrement exposé afin d'étudier la survenue d'un évènement dans un délai donné avec $Y = 1$ si le patient est en échec (ex : récurrence du cancer, progression tumorale, décès, ...), et $Y = 0$ dans le cas contraire. Comme expliqué précédemment, il n'est pas possible pour un patient i d'avoir à la fois une mesure de $Y_i^{(0)}$ et $Y_i^{(1)}$ du fait du caractère évolutif de la maladie en cancérologie ; cela est notamment connu comme le « problème fondamental de l'inférence causale » (Holland, 1986). Mais dans le cadre d'un essai clinique randomisé rendant les deux groupes de traitement comparables et homogènes avant administration du traitement, Rubin (2005) montre qu'il est possible de décomposer (2.1) et (2.2) comme :

$$\begin{aligned}
\delta(X) &= \mathbf{E}[Y^{(1)} - Y^{(0)}|X] \\
&= \mathbf{E}[Y^{(1)}|X] - \mathbf{E}[Y^{(0)}|X] \\
&= \mathrm{P}(Y = 1|Z = 1, X) - \mathrm{P}(Y = 1|Z = 0, X) \\
&= \rho_1(X) - \rho_0(X)\,,
\end{aligned}$$

ainsi que

$$\begin{aligned}
\Delta &= \mathbf{E}[Y^{(1)} - Y^{(0)}] \\
&= \mathbf{E}[Y^{(1)}] - \mathbf{E}[Y^{(0)}] \\
&= \mathrm{P}(Y = 1|Z = 1) - \mathrm{P}(Y = 1|Z = 0) \\
&= \rho_1 - \rho_0\,.
\end{aligned}$$

Il est possible de représenter graphiquement $\rho_1(X)$ et $\rho_0(X)$ en fonction des valeurs de X. Ce

graphique, appelé « marker-by-treatment predictiveness curves » dans la littérature par Janes et al. (2011) (nous les appellerons ici « courbes de risques »), permet d'avoir un premier aperçu des variations de $\delta(X)$, et de l'intérêt du marqueur X pour l'aide au choix de traitement.

Janes et al. (2011) proposent de représenter les courbes de risques prédits en ordonnée vs. la fonction de répartition du marqueur en abscisse afin de faciliter la comparaison entre plusieurs marqueurs d'échelles différentes.

La Figure 2.1 présente différents marqueurs simulés à partir de plusieurs modèles de régression logistique exprimés sous la forme :

$$\log\left[\frac{\mathrm{P}(Y=1|X,Z)}{\mathrm{P}(Y=0|X,Z)}\right] = \beta_0 + \beta_X \times X + \beta_Z \times Z + \beta_I \times XZ\,.$$

La justification de l'utilisation de ce modèle est donnée dans la section 2.2. Ces différents marqueurs simulés permettent d'illustrer les définitions énoncées précédemment dans la section 1.2.

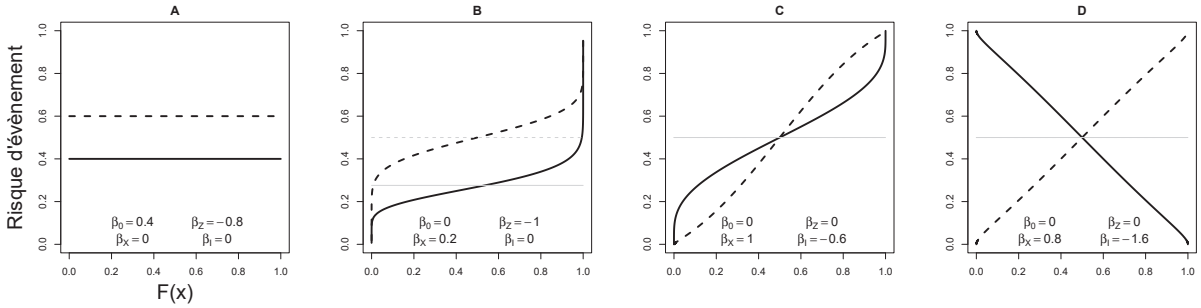


FIGURE 2.1 – Exemples de courbes de risque pour quatre marqueurs simulés
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence ; En gris : risques moyens de survenue d'évènement

Sur cette figure, les courbes de risques associées à quatre marqueurs (A, B, C et D) sont représentées. Pour le marqueur A, le risque de survenue de l'évènement étudié n'évolue pas en fonction des valeurs du marqueur, et la différence de risques entre les deux bras de traitement reste constante ; ce marqueur n'a pas de capacité prédictive. Pour le marqueur B, le risque de survenue de l'évènement évolue dans chaque bras de traitement avec les valeurs du marqueur. En effet, plus les valeurs du marqueur sont élevées, plus le risque de survenue de l'évènement augmente. Cependant, la différence de risques entre les deux bras de traitement reste constante. Il s'agit donc ici d'un marqueur pronostique, mais qui ne possède pas de caractère prédictif. Pour le marqueur C, le risque de survenue de l'évènement augmente dans les deux bras de traitement. Mais, dans ce cas, la différence de risques entre les deux bras n'est pas constante en fonction des valeurs du marqueur. Le marqueur C est donc un marqueur prédictif. Enfin, on obtient la même conclusion avec le marqueur D pour lequel il existe des variations dans la différence de risques encore plus importantes. Le fait que les variations dans la différence de risques soient plus importantes permet de mieux discriminer les catégories de patients qui retirent un bénéfice du traitement de référence de ceux qui retirent un bénéfice du traitement

innovant.

Dans le Chapitre 3, des méthodes permettant d'estimer le seuil optimal d'un marqueur (tenant compte à la fois de l'efficacité et des effets indésirables des deux traitements) seront présentées. Si les toxicités des deux traitements étaient négligeables, il serait possible de définir le seuil optimal du marqueur comme la valeur de marqueur pour laquelle les courbes de risques se croisent. Par exemple, pour les marqueurs C et D présentés dans la Figure 2.1, la valeur de seuil optimal serait la valeur de marqueur correspondant au deuxième quartile de la fonction de répartition du marqueur. La plupart des méthodes présentées par la suite feront l'hypothèse qu'il existe une valeur seuil de marqueur unique, cela se traduira sous l'hypothèse que $\delta(X)$ évolue de manière monotone en fonction de X.

Cette illustration vient appuyer la définition que l'on donnait précédemment d'un marqueur prédictif et semble également indiquer que la différence de risques est la mesure clé dans l'évaluation de tels marqueurs.

2.1.2 La recherche du marqueur parfait

Les différentes métriques qui seront présentées par la suite souffrent parfois d'une difficulté à être interprétées convenablement. La définition du marqueur « parfait » ou « idéal » est alors justifiée dans le but de pouvoir calculer la valeur maximale atteignable par ces métriques pour une situation donnée (Huang et al., 2015).

Le marqueur parfait est défini comme étant un marqueur qui permettra de recommander le traitement de référence à tous les patients qui retirent un bénéfice du traitement de référence, et de recommander le traitement innovant à tous les patients qui retirent un bénéfice du traitement innovant. Il est à noter que certains patients ne retireront un bénéfice ni du traitement de référence, ni du traitement innovant. Pour ces patients, peu importe ce que préconise le marqueur, puisque cela n'aura pas de conséquences. Le marqueur parfait est donc un marqueur qui permet d'éviter la survenue de l'évènement étudié chez tous les patients pour lesquels l'évènement est évitable avec l'un ou l'autre des traitements étudiés.

Lors de l'évaluation d'un critère de jugement binaire, il est possible d'identifier quatre situations pour un patient donné (Huang et al., 2015). Le Tableau 2.1 présente ces différentes situations.

Tableau 2.1 – Situations possibles selon la réponse à chacun des traitements

$Y^{(0)}$	$Y^{(1)}$	Situation	Proportion
1	0	Bénéfice du traitement innovant	q_1
1	1	Réponse identique	q_2
0	0	Réponse identique	q_3
0	1	Bénéfice du traitement de référence	q_4

A partir des informations fournies par ce tableau, il est possible de représenter schématiquement ce à quoi peuvent ressembler les courbes de risque associées au marqueur parfait, en fonction des quantités q_1, q_2, q_3 et q_4 (Figure 2.2).

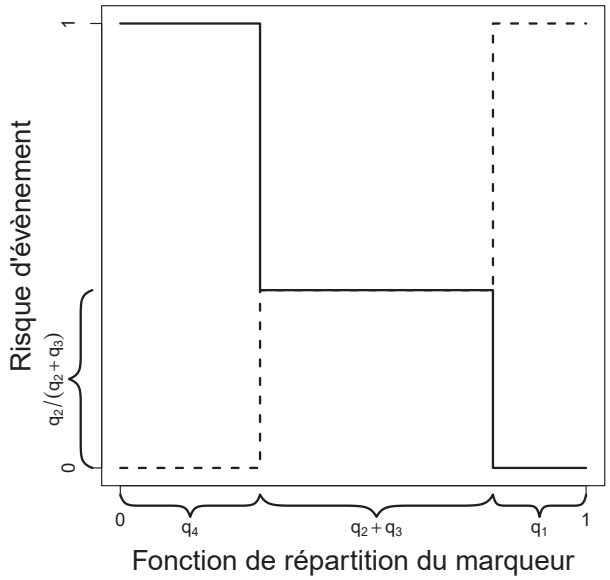


FIGURE 2.2 – Exemple de courbes de risque associées au marqueur parfait en fonction des valeurs de q_1, q_2, q_3 et q_4
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

Comme expliqué précédemment dans la section 1.2, il n'est pas possible de connaître la réponse de chaque patient aux deux traitements puisque chaque patient n'est randomisé que dans un seul bras de traitement. Cette contrainte rend non identifiables les quantités q_1, q_2, q_3 et q_4. Cependant, il est possible d'écrire les relations suivantes :

$$\begin{aligned} &q_1 + q_2 = \rho_0 \Rightarrow q_1 \leq \rho_0 \,, \\ &q_1 + q_3 = 1 - \rho_1 \Rightarrow q_1 \leq 1 - \rho_1 \,, \\ &q_1 - q_4 = \rho_0 - \rho_1 \Rightarrow q_1 \geq \rho_0 - \rho_1 \,, \end{aligned} \tag{2.3}$$

ce qui implique que $\max(0, \rho_0 - \rho_1) \leq q_1 \leq \min(\rho_0, 1 - \rho_1)$. Comme ρ_0 et ρ_1 sont identifiables, il suffit de fixer q_1 pour déterminer q_2, q_3, et q_4. Les courbes de risque associées au marqueur parfait sont donc fonction des risques moyens dans chaque bras (ρ_0 et ρ_1), mais pas seulement.

Prenons l'exemple d'une étude pour laquelle on sait que $\rho_0 = 0.5$ et $\rho_1 = 0.4$. Les inégalités présentées dans le système (2.3), permettent de déterminer les bornes de q_1 : $0.1 \leq q_1 \leq 0.5$. En faisant varier q_1 sur cet intervalle, il est possible de représenter les courbes de risque associées au marqueur parfait. Ces courbes de risque sont présentées dans la Figure 2.3.

Il est impossible en pratique de savoir lequel de ces graphiques correspond au vrai marqueur parfait d'une étude donnée. Une approche conservatrice serait de retenir le dernier graphique comme définition d'un marqueur parfait. En effet, pour ce graphique il est fait l'hypothèse que $q_1 = \min(\rho_0, 1 - \rho_1)$; sous cette hypothèse soit q_2 soit q_3 est nul. En réalité, en fonction des traitements considérés, il n'est pas sûr que cette situation soit réellement atteignable, et ce quel que soit le marqueur considéré. Une métrique qui évaluera les performances prédictives d'un marqueur relativement aux performances de ce marqueur parfait risquera dans certains cas de

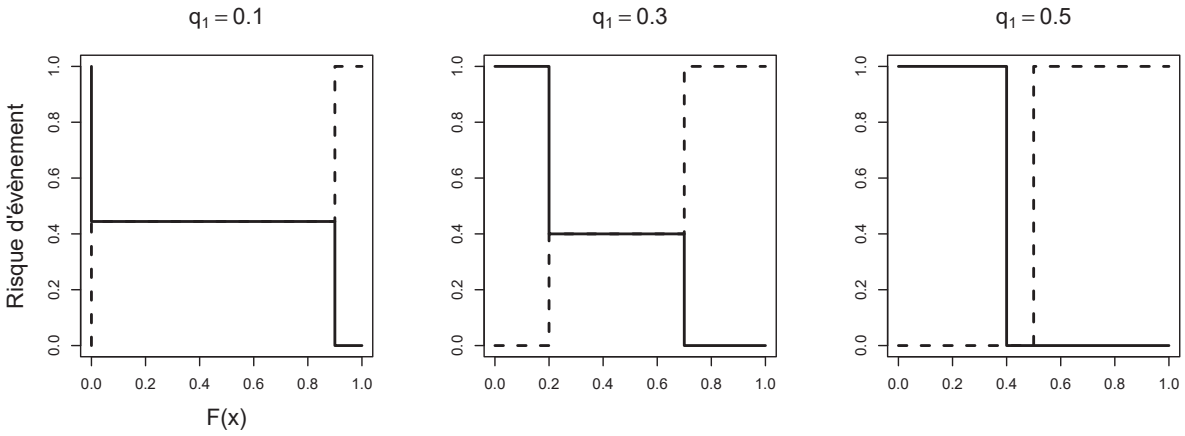


FIGURE 2.3 – Exemples de courbes de risque associées au marqueur parfait en fonction des valeurs de q_1
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

sous-estimer les performances des marqueurs étudiés (c'est pour cette raison que l'on parle d'une approche conservatrice).

On notera également que sous cette hypothèse, un marqueur prédictif parfait est un marqueur reflétant parfaitement les mécanismes d'action de chaque traitement et dont les valeurs basses sont associées à l'efficacité d'un traitement donné, tandis que les valeurs hautes sont associées à l'efficacité de l'autre traitement.

Pour la suite, nous retiendrons cette hypothèse pour caractériser un marqueur prédictif parfait.

2.2 Recherche d'une interaction entre le marqueur et les traitements

Au vu de la définition d'un marqueur prédictif (c'est-à-dire un marqueur pour lequel la différence de risques entre les deux traitements varie selon les valeurs du marqueur), la première idée qui vient à l'esprit pour détecter un marqueur prédictif est de modéliser le risque de survenue de l'évènement observé avec l'introduction d'une interaction entre le marqueur et les traitements étudiés. C'est d'ailleurs cette approche qui est proposée par Byar (1985) afin d'identifier de possibles marqueurs prédictifs au cours de l'analyse d'un essai clinique. Le modèle est défini par :

$$g[\mathbf{E}(Y|X,Z)] = \beta_0 + \beta_X \times X + \beta_Z \times Z + \beta_I \times XZ, \tag{2.4}$$

avec $g(.)$ la fonction de lien, $\mathbf{E}(Y|X,Z)$ l'espérance conditionnelle de la variable Y sachant X et Z, et $\boldsymbol{\beta} = (\beta_0, \beta_X, \beta_Z, \beta_I)'$ le vecteur de paramètres à estimer.

Dans son approche, l'auteur distingue deux types d'interactions, les interactions « quantitatives » et « qualitatives ». Pour l'interaction quantitative, la différence de risques entre les deux bras de traitement varie avec les valeurs du marqueur, mais le signe de cette différence ne varie pas. Pour l'interaction qualitative, la différence de risques varie également avec les valeurs

du marqueur, et le signe de cette différence est amené à changer. Ces deux types d'interaction sont présentés dans la Figure 2.4 avec en Figure A l'interaction quantitative et en Figure B l'interaction qualitative.

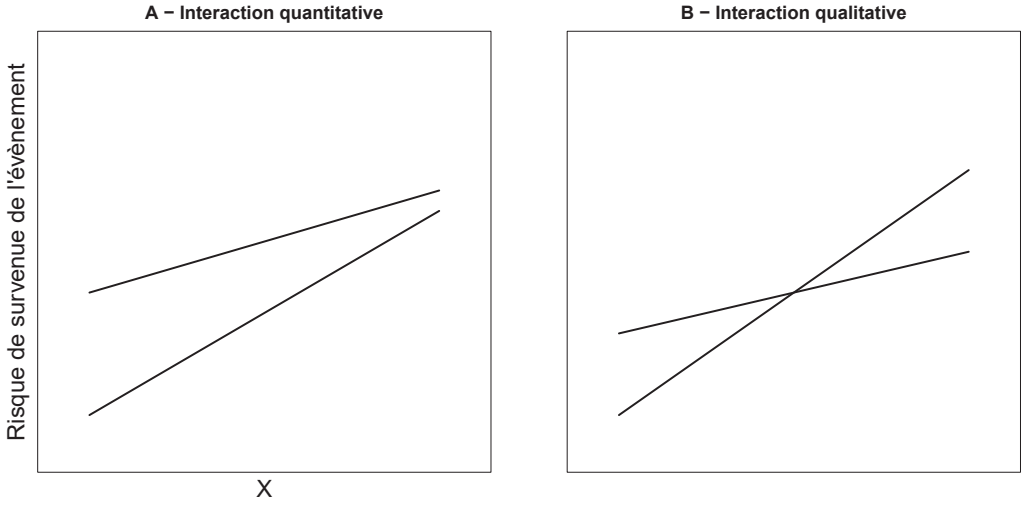


FIGURE 2.4 – Deux types d'interaction marqueur-traitement

A première vue, le type d'interaction recherché pour les marqueurs prédictifs est l'interaction qualitative puisque cela signifie que l'effet du traitement s'inverse selon les valeurs du marqueur et donc qu'un sous-groupe de patients retirera un bénéfice plus important d'une thérapie que de l'autre. Cependant, dans le cas où le critère de jugement étudié ne permet pas de tenir compte des toxicités des deux traitements, alors une interaction quantitative peut trouver son intérêt. Cela signifie que même si un traitement est moins efficace, il pourrait être préférable de le recommander pour des valeurs de marqueur pour lesquelles l'écart de risque entre les traitements est faible s'il est moins toxique, ou moins invasif, que l'autre traitement. Les notions d'utilité et de toxicité des traitements sont abordées dans le Chapitre 3.

La détection d'une interaction quantitative est dépendante du modèle utilisé (par exemple, une interaction dans un modèle additif peut ne pas exister dans un modèle multiplicatif). En revanche, les interactions qualitatives peuvent être détectées à la fois avec un modèle additif et un modèle multiplicatif (bien que l'amplitude du coefficient d'interaction, ainsi que son niveau de significativité diffèreront d'un modèle à l'autre).

Bien qu'une interaction puisse être détectée graphiquement, le recours à un test de Wald est nécessaire pour tester si le coefficient d'interaction du modèle est égal à 0.

Cette méthode est simple d'utilisation, et il s'agit de la méthode la plus couramment utilisée pour identifier des marqueurs prédictifs. La valeur du coefficient d'interaction donne une information quant à la capacité prédictive du marqueur, cependant son amplitude varie selon le modèle utilisé (additif ou multiplicatif), et selon l'échelle du marqueur étudié, rendant la comparaison entre différents marqueurs difficile. Cette approche ne permet pas non plus de prendre en compte le coût clinique de l'utilisation du marqueur en pratique.

De plus, au regard de la définition donnée d'un marqueur prédictif, il semblerait que seul

le coefficient d'interaction dans un modèle additif pourrait faire sens. Cependant, la plupart des modèles de risque développés sont des modèles multiplicatifs, à la fois pour des raisons conceptuelles et des raisons techniques (le risque prédit doit être borné entre 0 et 1). C'est le cas du modèle logistique, qui est classiquement retenu pour modéliser des risques à délai fixé en l'absence de censure (Janes et al., 2014a). Par ailleurs, comme tout modèle, le modèle logistique fait un certain nombre d'hypothèses. Dans l'équation (2.4), en prenant comme fonction de lien $g() = \text{logit}()$, il est fait l'hypothèse que l'effet du marqueur est linéaire sur l'échelle logit du risque, ainsi que l'interaction avec le traitement. Il est possible d'assouplir cette hypothèse, en ayant recours à des polynômes et des « splines », mais cela rend l'interprétation du test d'interaction complexe.

Toutes les limites énoncées précédemment concernant le test d'interaction ont poussé la communauté scientifique à développer de nouvelles métriques permettant de quantifier la capacité prédictive d'un marqueur.

2.3 Approches directes de la mesure du bénéfice moyen

Comme énoncé dans la section 1.2, il serait souhaitable de pouvoir prédire le différentiel de réponse aux deux traitements de manière individuelle. Cependant, cela est impossible dans la plupart des essais (particulièrement en oncologie où les essais en cross-over ne sont pas envisageables du fait du caractère évolutif de la maladie) car chaque patient ne reçoit qu'un seul des deux traitements et qu'il est donc impossible de mesurer directement le bénéfice individuel.

Cependant, plusieurs méthodes ont été proposées en s'appuyant sur les travaux menés dans le cadre de l'inférence causale (Rubin, 2005), et sont présentées dans cette section.

2.3.1 Approche par courbe ROC du bénéfice individuel

L'objectif d'un marqueur prédictif est de permettre l'identification de sous-groupes de patients qui retirent plutôt un bénéfice de l'un des deux traitements, tandis que d'autres sous-groupes retirent un bénéfice du deuxième traitement. Dans cette optique, cela revient à discriminer les patients de l'essai en les catégorisant soit dans un groupe qui retire un bénéfice du traitement innovant, soit dans un groupe qui retire un bénéfice du traitement de référence. Une telle problématique fait naturellement penser à une approche par courbe ROC, et c'est d'ailleurs ce qui a été proposé par Huang et al. (2012).

Pour rappel, le Tableau 2.1 permettait de définir quatre classes de patients en se basant sur les réponses individuelles à chacun des traitements. Comme expliqué précédemment, il n'est pas possible de savoir à quelle classe appartient un patient donné car on ne connaît pas les réponses individuelles à chaque traitement. Les auteurs effectuent donc plusieurs hypothèses constituant le « potential outcomes framework », ainsi qu'une hypothèse supplémentaire qui permettra d'estimer la bénéfice individuel. Les quatre hypothèses sont décrites ci-après :

Hypothèse (A-2.1). *L'unité de traitement est stable et consistante (SUTVA).*

Cette hypothèse signifie que pour un même patient $Y^{(0)}$ et $Y^{(1)}$ sont indépendants de l'allocation de traitement des autres patients, et $Y = Y^{(Z)}$ le critère de jugement réellement observé chez les patients.

Hypothèse (A-2.2). *L'allocation de traitement peut être ignorée :* $(Y^{(0)}, Y^{(1)}, X) \perp Z$.

Cette hypothèse signifie que la valeur de marqueur X ne dépend pas du fait que le patient soit traité avec le traitement de référence ou innovant, et que le critère de jugement potentiel ne dépend pas non plus du fait que le patient ait réellement été traité avec le traitement de référence ou innovant.

Hypothèse (A-2.3). *L'effet du traitement est monotone :* $Y^{(0)} \geq Y^{(1)}$.

Cette hypothèse signifie que le traitement innovant sera systématiquement au moins aussi efficace que le traitement de référence pour tous les patients.

Hypothèse (A-2.4). $\mathbf{E}[Y^{(Z)}|X]$ *peut être modélisée grâce un modèle linéaire généralisé.*

L'hypothèse (A-2.3) est l'hypothèse la plus forte, et c'est celle qui va permettre de quantifier le nombre de patients dans chaque classe. Cette hypothèse revient à dire que la proportion q_4 est égale à 0, et donc qu'aucun patient ne retire un bénéfice du traitement de référence car le traitement innovant est toujours au moins aussi efficace que le traitement de référence. Typiquement, ce genre d'hypothèse pourrait s'envisager dans la comparaison d'une nouvelle thérapie à un placebo, en supposant toutefois que cette thérapie n'ait aucun effet délétère. Fixer la proportion de patients dans l'une des quatre classes rend identifiables les proportions de patients dans chacune des autres classes en suivant le même principe que celui présenté dans la section 2.1.

Huang et al. (2012) proposent une approche par courbe ROC dans le but de discriminer les patients qui retirent un bénéfice du traitement innovant de ceux qui n'en retirent aucun.

Notations

On définit $B = \mathbb{1}(Y^{(0)} > Y^{(1)})$ l'indicatrice permettant de savoir si le patient retire un bénéfice de la stratégie thérapeutique innovante ($B = 1$) ou bien si le traitement innovant n'est pas plus efficace que le traitement de référence ($B = 0$). Dans l'exemple qui suit, on considèrera que des valeurs élevées de X sont synonymes d'un meilleur bénéfice du traitement innovant. Il est alors possible d'écrire :

$$\begin{aligned}
\mathbf{E}(B|X) &= \mathrm{P}(Y^{(0)} = 1, Y^{(1)} = 0|X) \\
&= \mathrm{P}(Y^{(0)} = 1|X) - \mathrm{P}(Y^{(0)} = 1, Y^{(1)} = 1|X) \\
&= \mathrm{P}(Y^{(0)} = 1|X) - \mathrm{P}(Y^{(0)} = 1|Y^{(1)} = 1, X)\mathrm{P}(Y^{(1)} = 1|X) \\
&= \mathrm{P}(Y^{(0)} = 1|X) - [1 - \mathrm{P}(Y^{(0)} = 0|Y^{(1)} = 1, X)]\mathrm{P}(Y^{(1)} = 1|X) \\
&= \mathrm{P}(Y^{(0)} = 1|X) - \mathrm{P}(Y^{(1)} = 1|X) \\
&= \rho_0(X) - \rho_1(X)\,,
\end{aligned}$$

où le résultat est obtenu grâce à l'hypothèse (A-2.3) qui permet de dire que $\mathrm{P}(Y^{(0)} = 0 | Y^{(1)} = 1, X) = 0$. Le principe d'une courbe ROC est de tracer les couples $(\mathrm{Se}(c),\ 1 - \mathrm{Sp}(c))$ pour l'ensemble des seuils c possibles. On complète les notations définies précédemment en définissant la sensibilité $(\mathrm{Se}(c))$ comme :

$$\begin{aligned}
\mathrm{Se}(c) &= \mathrm{P}(X > c | B = 1) \\
&= \frac{\mathrm{P}(B = 1, X > c)}{\mathrm{P}(B = 1)} \\
&= \frac{\mathbf{E}[B \times \mathbb{1}(X > c)]}{\mathbf{E}(B)} \\
&= \frac{\mathbf{E}[\mathbf{E}(B|X) \times \mathbb{1}(X > c)]}{\mathbf{E}[\mathbf{E}(B|X)]} \\
&= \frac{\mathbf{E}\{[\rho_0(X) - \rho_1(X)] \times \mathbb{1}(X > c)\}}{\mathbf{E}[\rho_0(X) - \rho_1(X)]}.
\end{aligned}$$

Sur le même principe, on peut exprimer le complément de la spécificité $(1 - \mathrm{Sp}(c))$ sous la forme suivante :

$$\begin{aligned}
1 - \mathrm{Sp}(c) &= \mathrm{P}(X > c | B = 0) \\
&= \frac{\mathbf{E}\{[1 - \rho_0(X) + \rho_1(X)] \times \mathbb{1}(X > c)\}}{\mathbf{E}[1 - \rho_0(X) + \rho_1(X)]}.
\end{aligned}$$

Pour pouvoir estimer la sensibilité et la spécificité, il est donc nécessaire de calculer $\rho_0(X)$ et $\rho_1(X)$. Pour ce faire les auteurs proposent d'utiliser un modèle logistique.

Le modèle de risque

Les hypothèses formulées par Huang et al. (2012) font que la sensibilité et la spécificité sont estimables à partir des paramètres d'un modèle de régression logistique modélisant simplement le risque de survenue de l'évènement étudié (donnée observée dans les essais) en fonction du traitement, du marqueur et de l'interaction marqueur-traitement :

$$\log\left[\frac{\mathrm{P}(Y = 1|X, Z)}{\mathrm{P}(Y = 0|X, Z)}\right] = \beta_0 + \beta_X \times X + \beta_Z \times Z + \beta_I \times XZ. \tag{2.5}$$

On note que sous l'hypothèse (A-2.3), $\rho_0(X) \geq \rho_1(X) \Leftrightarrow \delta(X) \leq 0$ pour tout X. Cependant, lors de l'estimation des paramètres du modèle de régression logistique (2.5), rien ne garantit que cette contrainte sera respectée. Pour qu'elle le soit, les auteurs proposent de maximiser la vraisemblance du modèle logistique (notée L) en imposant la contrainte de l'hypothèse (A-2.3). On définit la log-vraisemblance du modèle de la manière suivante :

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log[\mathrm{P}(Y_i = 1|X_i, Z_i)] + (1 - Y_i) \log[\mathrm{P}(Y_i = 0|X_i, Z_i)].$$

La maximisation de cette log-vraisemblance sous les contraintes définies peut se faire à l'aide d'un algorithme d'optimisation de type Nelder-Mead (Nelder and Mead, 1965).

Les estimateurs de la sensibilité et de la spécificité sont alors calculés naturellement en utilisant les estimateurs de $\rho_0(X)$ et de $\rho_1(X)$ (notés respectivement $\widehat{\rho}_0(X)$ et $\widehat{\rho}_1(X)$) obtenus grâce au modèle de régression logistique (2.5) :

$$\widehat{\mathrm{Se}}(c) = \frac{\int_c^{+\infty} \widehat{\rho}_0(x) - \widehat{\rho}_1(x)\, d\widehat{F}(x)}{\int_{-\infty}^{+\infty} \widehat{\rho}_0(x) - \widehat{\rho}_1(x)\, d\widehat{F}(x)},$$
$$1 - \widehat{\mathrm{Sp}}(c) = \frac{\int_c^{+\infty} 1 - \widehat{\rho}_0(x) + \widehat{\rho}_1(x)\, d\widehat{F}(x)}{\int_{-\infty}^{+\infty} 1 - \widehat{\rho}_0(x) + \widehat{\rho}_1(x)\, d\widehat{F}(x)},$$

où $\widehat{F}(.)$ est la fonction de répartition empirique du marqueur X.

Compléments et illustration de l'approche

L'aire sous la courbe ROC de Huang (notée θ_H) indique la capacité du marqueur à discriminer les patients qui tirent un bénéfice du traitement innovant de ceux qui n'en tirent aucun. Etant donné que la sensibilité et la spécificité sont bornées entre 0 et 1, θ_H est également bornée entre 0 et 1. Plus θ_H est proche de 1, plus le pouvoir discriminant du marqueur est important (et donc plus sa capacité prédictive est importante également). Lorsque θ_H se rapproche de 0.5, le marqueur n'a aucune capacité prédictive. Ce point, non cité par les auteurs, peut être démontré de la manière suivante.

Lorsqu'un modèle de risque est utilisé pour évaluer la capacité diagnostique (ou pronostique) d'un marqueur, il a été démontré qu'une correspondance existait entre la courbe de risque et l'aire sous la courbe ROC liée au marqueur étudié (Viallon and Latouche, 2011) :

$$\theta = \frac{\int_{-\infty}^{+\infty} F(x)\rho(x)\, dF(x) - \frac{\rho^2}{2}}{\rho(1-\rho)},$$

où θ est l'aire sous la courbe ROC pour un marqueur diagnostique (ou pronostique), $\rho(X) = \mathrm{P}(Y=1|X)$, et $\rho = \mathrm{P}(Y=1)$. Dans le cas des marqueurs prédictifs, il est donc possible d'étendre ces démonstrations et d'exprimer θ_H comme une fonction des courbes de risque dans chaque bras de traitement :

$$\theta_H = \frac{\int_{-\infty}^{+\infty} F(x)[\rho_0(x) - \rho_1(x)]dF(x) - \frac{(\rho_0-\rho_1)^2}{2}}{(\rho_0-\rho_1)(1-\rho_0+\rho_1)}. \tag{2.6}$$

Si la différence de risques entre les deux bras de traitement est constante alors la formule se récrit de la manière suivante :

$$\begin{aligned}\theta_H &= \frac{\frac{\rho_0-\rho_1}{2} - \frac{(\rho_0-\rho_1)^2}{2}}{(\rho_0-\rho_1)(1-\rho_0+\rho_1)} \\ &= 0.5 \times \frac{(\rho_0-\rho_1) - (\rho_0-\rho_1)^2}{(\rho_0-\rho_1) - (\rho_0-\rho_1)^2}\end{aligned}$$

$$= 0.5\,.$$

S'il n'existe aucune variation de la différence de risques entre les deux bras, en fonction des valeurs du marqueur, alors le marqueur n'a aucun intérêt prédictif selon cette méthode. Cette logique va dans le sens de la définition donnée précédemment d'un marqueur prédictif. *A contrario*, plus les variations de $\delta(X)$ sont importantes, plus le marqueur a une valeur de θ_H élevée.

Dans l'équation (2.6), les variations de $\delta(X)$ sont évaluées dans l'intégrale. On constate que plus la différence de risques marginale est faible, plus de faibles variations de $\delta(X)$ sont suffisantes pour avoir une valeur de θ_H élevée. *A contrario*, avec une différence de risques marginale élevée, il faut de plus grandes variations de $\delta(X)$ pour atteindre une valeur de θ_H élevée.

On peut néanmoins nuancer l'interprétation qu'on donne à θ_H. En effet les auteurs se placent dans un cas très particulier fixé par l'hypothèse (A-2.3), puisque l'objectif n'est plus d'identifier un sous-groupe de patients bénéficiant plus du traitement innovant ou du traitement de référence, mais plutôt d'identifier un sous-groupe de patients pour lequel la thérapie innovante n'est pas plus efficace qu'un placebo afin d'éviter les effets indésirables potentiels d'une telle thérapie. Cette question peut donc trouver son intérêt dans certaines situations, mais elle est peu envisageable lorsque le traitement de référence n'est pas un placebo.

Huang et al. (2012) se limitent donc à l'évaluation de marqueurs pour lesquels une interaction quantitative existe (excluant de fait les interactions qualitatives). Il s'agit là de la principale limite de cette approche.

La Figure 2.5 présente les courbes de risque en fonction des valeurs du marqueur étudié permettant d'atteindre une valeur de θ_H de 1 avec la méthode proposée ci-dessus avec un risque moyen égal à 0.82 et 0.12 dans les bras référence et innovant, respectivement. On constate ici que la définition du marqueur parfait, sous l'hypothèse (A-2.3) n'est qu'un cas particulier de celle qui avait été retenue dans la sous-section 2.1.2.

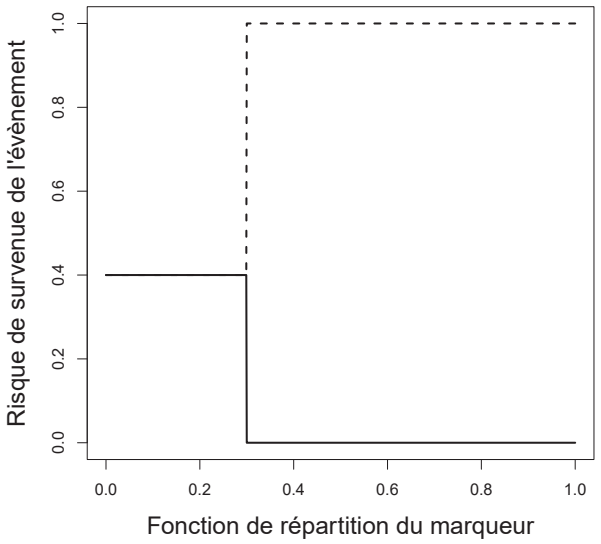


FIGURE 2.5 – Exemple de courbes de risque pour obtenir une valeur de θ_H égale à 1
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

2.3.2 Extension de l'approche par courbe ROC du bénéfice individuel

Zhang et al. (2014) ont proposé une extension de la méthode de Huang et al. (2012) afin de se passer de l'hypothèse (A-2.3) qui était posée par ces derniers. On note $B = \mathbb{1}(Y^{(0)} > Y^{(1)})$ qui prend comme valeur 1 si le patient a une meilleure réponse au traitement innovant qu'au traitement de référence et 0 dans le cas contraire. On peut également noter $B = \mathbb{1}(Y^{(0)} \geq Y^{(1)})$ si le traitement innovant est préférable au traitement de référence à efficacités comparables (notation qui n'était pas possible avec l'approche de Huang et al. (2012)).

Les expressions de la sensibilité et de spécificité s'en trouvent impactées car il n'est plus possible de dire que $\mathrm{P}(B = 1|X) = \rho_0(X) - \rho_1(X)$. On exprime donc la sensibilité et la spécificité sous les formes suivantes

$$\begin{aligned}
\mathrm{Se}(c) &= \frac{\mathbf{E}[\mathbf{E}(B|X) \times \mathbb{1}(X > c)]}{\mathbf{E}[\mathbf{E}(B|X)]} \\
&= \frac{\int_c^{+\infty} \mathrm{P}(B = 1|X = x)\, dF(x)}{\int_{-\infty}^{+\infty} \mathrm{P}(B = 1|X = x)\, dF(x)}, \\
1 - \mathrm{Sp}(c) &= \frac{\mathbf{E}\{[1 - \mathbf{E}(B|X)] \times \mathbb{1}(X > c)\}}{\mathbf{E}[1 - \mathbf{E}(B|X)]} \\
&= \frac{\int_c^{+\infty} 1 - \mathrm{P}(B = 1|X = x)\, dF(x)}{\int_{-\infty}^{+\infty} 1 - \mathrm{P}(B = 1|X = x)\, dF(x)}.
\end{aligned}$$

On notera par la suite $\pi_X(x) = \mathrm{P}(B = 1|X = x)$. Malgré son apparence simple, $\pi_X(x)$ n'est pas facile à estimer puisqu'il fait intervenir la distribution jointe de $Y^{(0)}$ et $Y^{(1)}$, et que la dépendance entre ces deux réponses n'est pas observable dans les données puisque chaque patient ne reçoit qu'un seul traitement. Les auteurs proposent donc de conditionner la distribution jointe de $Y^{(0)}$ et $Y^{(1)}$ sur le marqueur étudié, ainsi que sur un ensemble de covariables mesurées avant l'administration des traitements et liées aux réponses de chaque traitement (covariables pronostiques). On note $\boldsymbol{V}$ le vecteur de covariables ajoutées et $\boldsymbol{W} = (X, \boldsymbol{V})$. Il est alors possible de définir $\pi_W(\boldsymbol{w}) = \mathrm{P}(B = 1|\boldsymbol{W} = \boldsymbol{w})$, et d'écrire :

$$\pi_X(x) = \mathbf{E}[\pi_W(\boldsymbol{W})|X = x] = \int \pi_W(x, \boldsymbol{v})\, dF(\boldsymbol{v}|x).$$

$F(\boldsymbol{v}|x)$ peut être estimée de manière empirique, la difficulté réside dans l'estimation de $\pi_W(\boldsymbol{w})$. Zhang et al. (2014) font alors l'hypothèse suivante :

Hypothèse (A-2.5). *Les critères de jugement potentiels sont indépendants conditionnellement aux covariables* $\boldsymbol{W}$ *:* $Y^{(0)} \perp Y^{(1)}|\boldsymbol{W}$.

Cette hypothèse revient à dire que, conditionnellement aux valeurs du marqueur X et à un ensemble d'autres variables noté $\boldsymbol{V}$, le devenir du patient avec le traitement innovant est indépendant de son devenir sous le traitement de référence. Si cette hypothèse permet d'estimer la distribution jointe de $Y^{(0)}$ et $Y^{(1)}$ sans recours à l'hypothèse (A-2.3) très réductrice de Huang et al. (2012), elle n'en reste pas moins forte et difficile à vérifier. On notera que si $\boldsymbol{W}$ n'est pas suffisant pour expliquer la dépendance entre $Y^{(0)}$ et $Y^{(1)}$ alors le reste des développements n'est

plus valide. C'est pour cela que Zhang et al. (2014) proposent, pour chaque méthode d'estimation, d'évaluer cette hypothèse via une analyse de sensibilité. Celle-ci consiste à introduire une variable latente dans les formules permettant de tenir compte de la dépendance non expliquée entre $Y^{(0)}$ et $Y^{(1)}$:

$$Y^{(0)} \perp Y^{(1)} | (\boldsymbol{W}, U), \tag{2.7}$$

où U est la variable latente indépendante de $\boldsymbol{W}$ et distribuée selon une loi normale centrée-réduite. Cette hypothèse n'est pas suffisante pour estimer $\pi_W(\boldsymbol{w})$ car U n'est pas observée. Les auteurs proposent alors d'introduire la variable U dans les modèles présentés ci-après sous la forme d'un effet aléatoire et de réaliser une analyse de sensibilité sur la variance de cet effet aléatoire. Les analyses de sensibilité sont présentées rapidement pour chacune des approches proposées pour estimer $\pi_W(\boldsymbol{w})$: une approche directe et une approche indirecte.

Approche directe

Ce que les auteurs appellent « approche directe » est la modélisation du bénéfice individuel, B. On construit donc un modèle paramétrique pour $\mathrm{P}(B = 1|\boldsymbol{W})$ tel que

$$g[\mathbf{E}(B|\boldsymbol{W})] = \alpha_0 + \boldsymbol{\alpha}'_W \boldsymbol{W}, \tag{2.8}$$

où g est la fonction de lien retenue (logit ou probit), et $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}'_W)'$ est le vecteur des paramètres de régression du modèle.

Le vecteur de paramètres $\boldsymbol{\alpha}$ n'est pas estimable directement à partir des données. Supposons que l'hypothèse d'indépendance conditionnelle (A-2.5) soit valide. Pour pouvoir bien comprendre l'approche proposée par Zhang et al. (2014), on considère tout d'abord que $\boldsymbol{W}$ est discret et prend des valeurs dans $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_K\}$. Dans chaque strate définie par $\boldsymbol{W} = \boldsymbol{w}_k$, l'hypothèse (A-2.5) implique que $Y^{(0)}$ est indépendant de $Y^{(1)}$, ce qui signifie que la paire $\{Y_i^{(0)}, Y_i^{(1)}\}$ est distribuée de manière identique à la paire $\{Y_i^{(0)}, Y_j^{(1)}\}$. Si $Z_i = 0$ et $Z_j = 1$, alors la paire $\{Y_i^{(0)}, Y_j^{(1)}\}$ est observable directement dans les données comme étant la paire $\{Y_i, Y_j\}$, de telle sorte que $\pi_W(\boldsymbol{w}_k) = \mathbf{E}(B|\boldsymbol{W} = \boldsymbol{w}_k)$ peut être estimée par

$$\frac{1}{n_{0k} n_{1k}} \sum_{i \in S_{0k}} \sum_{j \in S_{1k}} B_{ij}, \tag{2.9}$$

où $B_{ij} = \mathbb{1}(Y_i > Y_j)$ ou $\mathbb{1}(Y_i \geq Y_j)$ selon la définition du bénéfice qui a été retenue, S_{0k} l'ensemble des patients dans le bras référence faisant partie de la strate $\boldsymbol{W} = \boldsymbol{w}_k$, S_{1k} l'ensemble des patients dans le bras innovant faisant partie de la strate $\boldsymbol{W} = \boldsymbol{w}_k$, et n_{zk} la taille de S_{zk} avec z prenant ses valeurs dans $\{0, 1\}$.

L'objectif est de généraliser ce concept aux marqueurs continus. Pour conserver la même approche que dans la formule (2.9), il faudrait trouver des patients dans chaque bras de traitement ayant les mêmes valeurs de $\boldsymbol{W}$, ce qui devient difficile dans le cas de marqueurs continus. Zhang et al. (2014) proposent donc d'estimer $\boldsymbol{\alpha}$ à l'aide d'un modèle de régression semi-paramétrique (Dodd and Pepe, 2003) qui servira de modèle intermédiaire dans l'estimation des paramètres

du modèle de bénéfice :

$$g[\mathbf{E}(B_{ij}|\boldsymbol{W}_i,\boldsymbol{W}_j)] = \beta_0 + \boldsymbol{\beta}'_W(\boldsymbol{W}_i+\boldsymbol{W}_j)/2 + \boldsymbol{\beta}'_{dW}(\boldsymbol{W}_j-\boldsymbol{W}_i)\,, \tag{2.10}$$

où $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}'_W, \boldsymbol{\beta}'_{dW})'$, S_z est la strate de patients pour lesquels $Z = z$ avec $i \in S_0$ et $j \in S_1$.

Lorsque l'hypothèse d'indépendance conditionnelle est respectée, on note que $\mathrm{P}(B = 1|\boldsymbol{W}_i = \boldsymbol{W}_j = \boldsymbol{w}) = \mathrm{P}(B = 1|\boldsymbol{W} = \boldsymbol{w})$. Alors le modèle (2.10) se réduit au modèle (2.8) avec

$$\boldsymbol{\alpha} = (\beta_0, \boldsymbol{\beta}'_W)'\,.$$

C'est pour cela que les auteurs parlent du modèle (2.10) comme une aide dans l'estimation des paramètres du modèle (2.8). Le vecteur de paramètres $\boldsymbol{\beta}$ peut alors être estimé par maximum de vraisemblance.

Comme suggéré par Dodd and Pepe (2003), il n'y a pas besoin d'inclure toutes les paires $\{i; j\}$ possibles dans le modèle (2.10), mais seulement celles pour lesquelles $||\boldsymbol{W}_i - \boldsymbol{W}_j|| < \varepsilon$ avec $||.||$ la norme euclidienne, et pour $\varepsilon > 0$. Le choix de ε revient à rechercher un compromis entre le biais dans l'estimation des paramètres, et la stabilité du modèle.

Lorsque l'hypothèse (A-2.5) n'est pas vérifiée, alors on ne peut pas écrire que $\boldsymbol{\alpha} = (\beta_0, \boldsymbol{\beta}'_W)'$. Zhang et al. (2014) proposent alors de modifier la modèle (2.10) en introduisant la variable latente U définie précédemment :

$$\begin{aligned} g[\mathbf{E}(B_{ij}|\boldsymbol{W}_i,\boldsymbol{W}_j,U_i,U_j)] = \zeta_0 &+ \boldsymbol{\zeta}'_W(\boldsymbol{W}_i+\boldsymbol{W}_j)/2 + \boldsymbol{\zeta}'_{dW}(\boldsymbol{W}_j-\boldsymbol{W}_i) \\ &+ \zeta_U(U_i+U_j)/2 + \zeta'_{dU}(U_j-U_i)\,. \end{aligned} \tag{2.11}$$

Il est alors possible d'écrire :

$$\boldsymbol{\alpha} = \gamma(\beta_0, \boldsymbol{\beta}'_W)'\,,$$

où γ est un scalaire calculé à partir des paramètres ζ_U et ζ_{dU} du modèle (2.11), il reflète donc les caractéristiques de la variable latente U définie dans la formule (2.7). Les auteurs ont démontré que les valeurs de γ étaient comprises dans l'intervalle $[2^{-1/2}; +\infty[$ (voir les informations supplémentaires en ligne de l'article de Zhang et al. (2014)). Lorsque l'hypothèse (A-2.5) a des raisons d'être remise en doute, il est alors possible de réaliser une analyse de sensibilité sur le paramètre γ, avec $\gamma = 1$ correspondant à l'hypothèse d'indépendance conditionnelle.

Approche indirecte

L'approche indirecte consiste à construire un modèle de risque incluant une interaction entre $\boldsymbol{W}$ et Z. On note le modèle de la manière suivante :

$$g[\mathbf{E}(Y|Z,\boldsymbol{W})] = \theta_0 + \theta_Z Z + \boldsymbol{\theta}'_W W + \boldsymbol{\theta}'_{ZW} Z\boldsymbol{W}\,, \tag{2.12}$$

où $\boldsymbol{\theta} = (\theta_0, \theta_Z, \boldsymbol{\theta}'_W, \boldsymbol{\theta}'_{ZW})'$ est le vecteur des paramètres de régression du modèle de risque et $g(.)$ est une fonction de lien de type logit ou probit.

Sous l'hypothèse d'indépendance conditionnelle (A-2.5), la distribution jointe de $Y^{(0)}$ et $Y^{(1)}$ est identifiée comme suit :

$$\begin{aligned}\mathrm{P}(Y^{(0)}=y_0,Y^{(1)}=y_1|\boldsymbol{W}=\boldsymbol{w}) &= f(y_0,y_1|\boldsymbol{w})\\ &= \mathrm{P}(Y=y_0|Z=0,\boldsymbol{W}=\boldsymbol{w})\mathrm{P}(Y=y_1|Z=1,\boldsymbol{W}=\boldsymbol{w}),\end{aligned}$$

et estimée à partir du modèle de risque défini précédemment.

Lorsque $B=\mathbb{1}(Y^{(0)}>Y^{(1)})$, on peut alors exprimer $\pi_W(\boldsymbol{w})$ de la manière suivante :

$$\begin{aligned}\pi_W(\boldsymbol{w}) &= \int\int \mathbb{1}(y_0>y_1) f(y_0,y_1|\boldsymbol{w})\, dy_0 dy_1\\ &= \mathrm{P}(Y=1|Z=0,\boldsymbol{W}=\boldsymbol{w})\mathrm{P}(Y=0|Z=1,\boldsymbol{W}=\boldsymbol{w}),\end{aligned}$$

où il est possible d'écrire la dernière ligne car le critère de jugement considéré dans cette thèse est un critère de jugement binaire. Cette quantité peut alors être estimée en utilisant les prédictions du modèle (2.12). Si le bénéfice avait été défini sous la forme $B=\mathbb{1}(Y^{(0)}\geq Y^{(1)})$ alors

$$\begin{aligned}\pi_W(\boldsymbol{w}) &= \mathrm{P}(Y=1|Z=0,\boldsymbol{W}=\boldsymbol{w})\mathrm{P}(Y=1|Z=1,\boldsymbol{W}=\boldsymbol{w})\\ &\quad+\mathrm{P}(Y=0|Z=0,\boldsymbol{W}=\boldsymbol{w})\mathrm{P}(Y=0|Z=1,\boldsymbol{W}=\boldsymbol{w})\\ &\quad+\mathrm{P}(Y=1|Z=0,\boldsymbol{W}=\boldsymbol{w})\mathrm{P}(Y=0|Z=1,\boldsymbol{W}=\boldsymbol{w}),\end{aligned}$$

pouvant également être estimée à partir des prédictions du modèle (2.12). Lorsque l'hypothèse (A-2.5) est remise en cause, les auteurs proposent de réaliser une analyse de sensibilité basée sur la formule (2.7) qui permet d'écrire :

$$\begin{aligned}\mathrm{P}(Y^{(0)}=&y_0,Y^{(1)}=y_1|\boldsymbol{W}=\boldsymbol{w}) = \int \mathrm{P}(Y^{(0)}=y_0,Y^{(1)}=y_1|\boldsymbol{W}=\boldsymbol{w},U=u)\mathrm{P}(U=u)\, du\\ &= \int \mathrm{P}(Y=y_0|Z=0,\boldsymbol{W}=\boldsymbol{w},U=u)\mathrm{P}(Y=y_1|Z=1,\boldsymbol{W}=\boldsymbol{w},U=u)\mathrm{P}(U=u)\, du.\end{aligned}$$

Cette écriture implique donc qu'un nouveau modèle de risque soit spécifié pour Y sachant $(Z,\boldsymbol{W},U)$. On définit donc le modèle linéaire généralisé mixte suivant :

$$g[\mathbf{E}(Y|Z,\boldsymbol{W},U)] = \theta_0^* + \theta_Z^* Z + \boldsymbol{\theta}_W^{*\prime} W + \boldsymbol{\theta}_{ZW}^{*\prime} Z\boldsymbol{W} + \theta_U^* U\,,$$

où $\boldsymbol{\theta}^* = (\theta_0^*, \theta_Z^*, \boldsymbol{\theta}_W^{*\prime}, \boldsymbol{\theta}_{ZW}^{*\prime}, \theta_U^*)'$, U est l'effet aléatoire inclus dans le modèle et θ_U^* représente l'écart-type de cet effet aléatoire. U représente ici un facteur pronostique non observé qui affecte les réponses aux deux traitements de la même manière. Etant donné que U n'est pas observé dans les données, le modèle défini n'est pas identifiable, les auteurs proposent donc de réaliser une analyse de sensibilité en faisant varier les valeurs de θ_U^* (Zhang et al., 2014).

Estimation de $\pi_X(x)$

Afin d'estimer $\pi_X(x)$ à partir des estimations de $\pi_W(\boldsymbol{w})$, les auteurs proposent d'utiliser une méthode semi-paramétrique. On définit $\xi : \boldsymbol{R} \to [0;+\infty[$ une fonction kernel et $\lambda > 0$ un paramètre de la taille de la fenêtre utilisée par la fonction. $\pi_X(x)$ est alors estimée par :

$$\widehat{\pi}_X(x) = \frac{\sum_{i=1}^{n} \xi\{(X_i - x)/\lambda\}\widehat{\pi}_W(x, \boldsymbol{V}_i)}{\sum_{i=1}^{n} \xi\{(X_i - x)/\lambda\}}.$$

On notera ici qu'un point important concerne le choix de λ pour lequel les auteurs proposent de réaliser une approche de validation croisée.

Il est alors possible d'estimer la sensibilité et la spécificité définies auparavant en injectant dans les formules les estimations nécessaires.

Compléments et illustration de l'approche

Bien que non défini par Zhang et al. (2014) dans leur article, il est possible d'exprimer l'aire sous la courbe ROC de Zhang (notée θ_{Zh}) en fonction de la quantité $\pi_X(x)$ en reprenant le principe proposé pour l'approche de Huang et al. (2012) dans la sous-section précédente :

$$\theta_{Zh} = \frac{\int_{-\infty}^{+\infty} F(x)\pi_X(x)\, dF(x) - \frac{\mathrm{P}(B=1)^2}{2}}{\mathrm{P}(B=1)[1-\mathrm{P}(B=1)]}.$$

En s'affranchissant de l'hypothèse (A-2.3) posée par Huang et al. (2012), Zhang et al. (2014) permettent l'étude d'interactions qualitatives en plus des interactions quantitatives dans la recherche d'un marqueur prédictif. Ceci impacte directement l'interprétation et le calcul de la courbe ROC puisque celle-ci aura désormais une AUC de 1 lorsque les variations de la différence de risques entre les deux bras de traitement sont maximales, ce qui correspond à la définition du marqueur prédictif « parfait » posée précédemment dans la sous-section 2.1.2. La Figure 2.6 illustre ces différentes situations.

Pour le marqueur présenté dans la colonne de gauche, on constate qu'un marqueur parfait, au sens défini par Huang et al. (2012) par leur hypothèse (A-2.3), ne permet pas d'atteindre une AUC de 1 avec la courbe ROC de Zhang. En revanche le marqueur présenté dans la colonne de droite est un marqueur parfait au sens qui a été défini dans la sous-section 2.1.2, et l'aire sous la courbe ROC de Zhang est ici égale à 1.

La Figure 2.1 présentait quatre marqueurs, les deux premiers étaient considérés comme non prédictifs, tandis que les deux derniers avaient une capacité prédictive à quantifier. L'approche de Zhang et al. (2014) permet de représenter les courbes ROC associées à ces quatre marqueurs afin de quantifier leur capacité prédictive respective. La Figure 2.7 décrit la correspondance entre les courbes de risque associées aux quatre marqueurs et les courbes ROC.

Pour le marqueur A, les courbes de risque étant toutes deux horizontales, il était possible d'affirmer que le marqueur n'avait aucune capacité prédictive. Cela se vérifie également avec la courbe ROC de Zhang pour laquelle $\theta_{Zh} = 0.5$. Pour le marqueur B, alors que le coeffi-

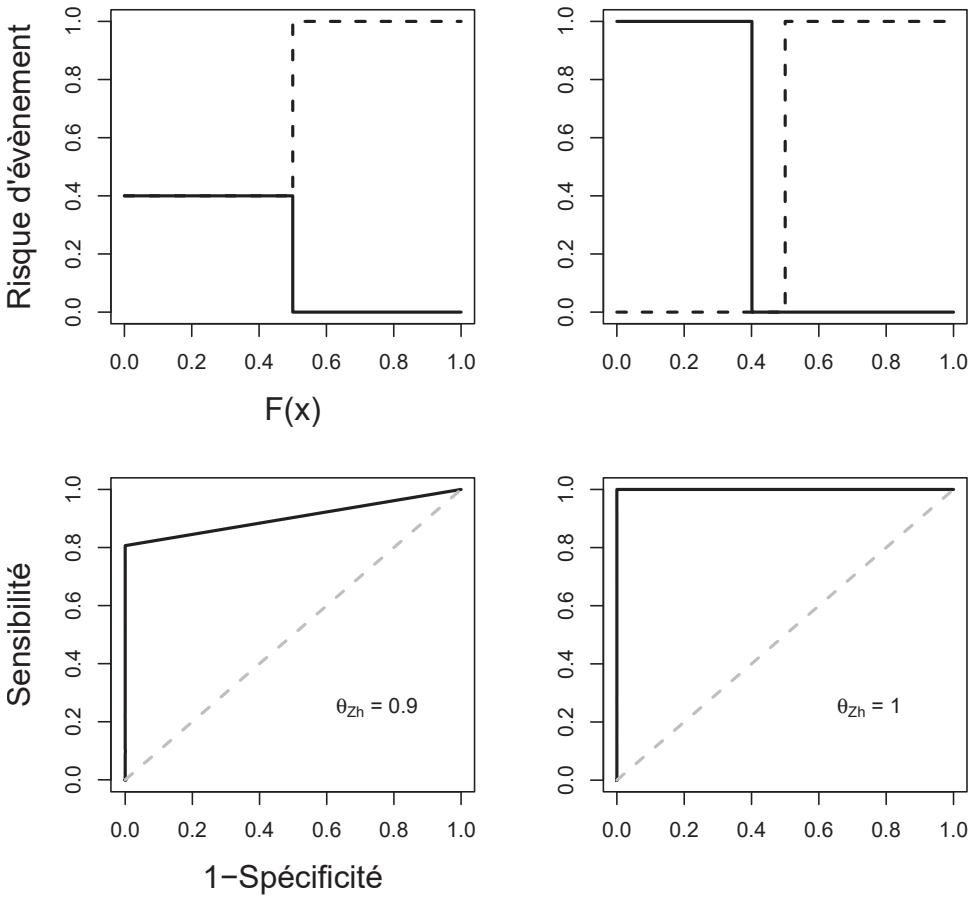


FIGURE 2.6 – Correspondance entre les courbes de risque d'un marqueur « parfait » et les courbes ROC de Zhang
Courbe pleine : Traitement innovant (graphiques du haut) ; Courbe en pointillés : Traitement de référence (graphiques du haut)

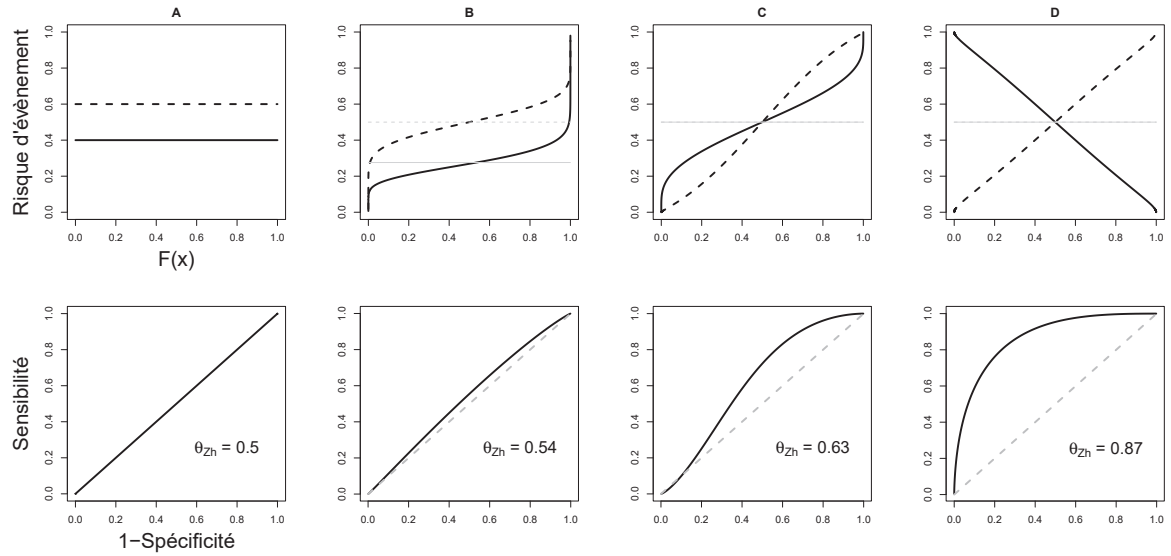


FIGURE 2.7 – Correspondance entre les courbes de risque de quatre marqueurs simulés et les courbes ROC de Zhang
Courbe pleine : Traitement innovant (graphiques du haut) ; Courbe en pointillés : Traitement de référence (graphiques du haut)

cient d'interaction du modèle logistique qui a permis de le simuler était égal à 0, on constate que $\theta_{Zh} = 0.54$. Selon l'approche de Zhang ce marqueur aurait donc une (faible) capacité prédictive. Ce phénomène est une illustration de la limite de l'approche du test d'interaction. En effet, l'interaction était nulle pour un modèle multiplicatif (modèle ayant permis de générer les courbes de risque), mais elle existe sur l'échelle additive, bien que faible. Ceci est reflété par la courbe ROC obtenue selon la méthode de Zhang et al. (2014). Concernant le marqueur C, la différence de risques variait avec les valeurs du marqueur, indiquant que celui-ci avait une capacité prédictive qui restait à quantifier. La courbe ROC de Zhang associée permet de confirmer cela et de quantifier $\theta_{Zh} = 0.63$. Enfin, le marqueur D était un marqueur pour lequel la différence de risque variait grandement avec les valeurs du marqueur, indiquant que celui-ci avait une capacité prédictive potentiellement plus importante que celle du marqueur C. Cela se vérifie avec la courbe ROC de Zhang associée car $\theta_{Zh} = 0.87$, bien supérieure à celle du marqueur C.

2.3.3 Bilan de ces approches

Face à la complexité imposée par la non-identifiabilité du bénéfice individuel, des auteurs ont tenté de proposer des approches s'appuyant sur les concepts de l'inférence causale (Rubin, 2005). Au-delà de ces concepts, Huang et al. (2012) ont été obligés de faire une hypothèse forte imposant que le traitement innovant soit toujours au moins aussi efficace que le traitement de référence. Cette hypothèse est grandement discutable, et ne pourrait se justifier que dans le cas de la comparaison d'un traitement innovant à un placebo (ce qui est rarement le cas en cancérologie). De plus, cette hypothèse conduit également à surestimer les capacités du marqueur dans sa capacité à guider le choix du traitement lorsqu'elle n'est pas justifiée, et à restreindre l'analyse aux interactions quantitatives.

Zhang et al. (2014) ont alors proposé une approche permettant de s'affranchir de cette hypothèse. Pour ce faire ils font une hypothèse d'indépendance des critères de jugement potentiels sachant un ensemble de covariables ayant un effet pronostique sur la survenue de l'évènement étudié. La principale limite ici est qu'il est nécessaire d'avoir recueilli l'ensemble de ces covariables expliquant la dépendance entre les critères de jugement potentiels, ce qui peut s'avérer assez difficile à vérifier en pratique. L'analyse de sensibilité qu'ils proposent permet d'avoir une idée de l'information expliquée par les covariables cependant elle peut rapidement rendre les résultats difficilement interprétables.

Toutes ces remarques ont par ailleurs également été relevées par Janes et al. (2015b), qui jugent ces hypothèses peu vraisemblables en pratique clinique.

2.4 Approches indirectes de la mesure du bénéfice moyen

Puisque l'évaluation du bénéfice individuel impose des contraintes fortes dans les méthodes précédemment exposées, d'autres auteurs ont préféré proposer des indicateurs permettant d'évaluer le caractère prédictif global d'un marqueur sans avoir à mesurer le bénéfice individuel.

2.4.1 Gain total

Le gain total (TG) est l'une des mesures globales proposées dans la littérature (Janes et al., 2014a). L'intérêt de cet indicateur est qu'il ne dépend pas d'une règle de décision basée sur un seuil. On l'exprime comme :

$$\mathrm{TG} = \int |\delta(X) - (\rho_1 - \rho_0)| \, dF_\delta \,,$$

où F_δ est la fonction de répartition de $\delta(X)$. Afin de comprendre ce que mesure le gain total, la Figure 2.8 représente $\delta(X)$ vs. F_δ pour un marqueur simulé.

Sur ces graphiques, on constate que la différence de risques varie de manière assez marquée selon les valeurs du marqueur. La ligne horizontale en gris correspond à la différence de risques marginale entre les deux bras de traitement. Si le marqueur n'avait aucun intérêt, les courbes noires et grises seraient superposées. Le gain total mesure donc l'aire entre la courbe en noire et la droite grise.

Pour rappel, la Figure 2.1 présentait quatre marqueurs, les deux premiers étaient considérés comme non prédictifs, tandis que les deux derniers avaient une capacité prédictive à quantifier. La Figure 2.9 permet de visualiser la correspondance entre les courbes de risque associées aux quatre marqueurs et le TG.

Pour le marqueur A, les courbes de risque étant toutes deux horizontales, il était possible d'affirmer que le marqueur n'avait aucune capacité prédictive. Cela se vérifie également avec le TG égal à 0. Pour le marqueur B, le TG restait très faible et égal à 0.02. Concernant le marqueur C, la différence de risques variait avec les valeurs du marqueur, indiquant que celui-ci avait une capacité prédictive qui restait à quantifier. Le TG associé permet de confirmer cela car il est

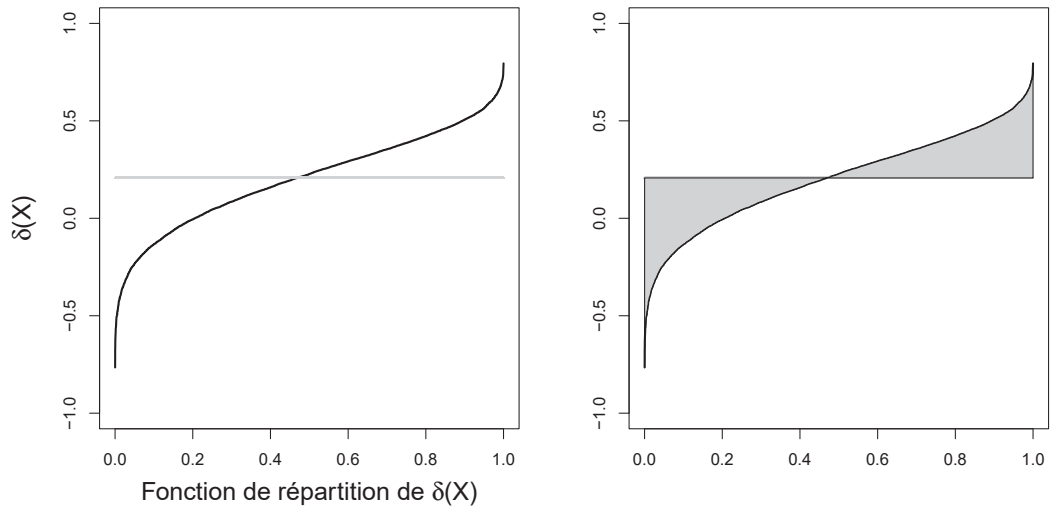


FIGURE 2.8 – Représentation de $\delta(X)$ en fonction de F_δ
Courbe pleine noire : $\delta(X)$; Courbe pleine grise : $\rho_1 - \rho_0$; Aire grisée : TG

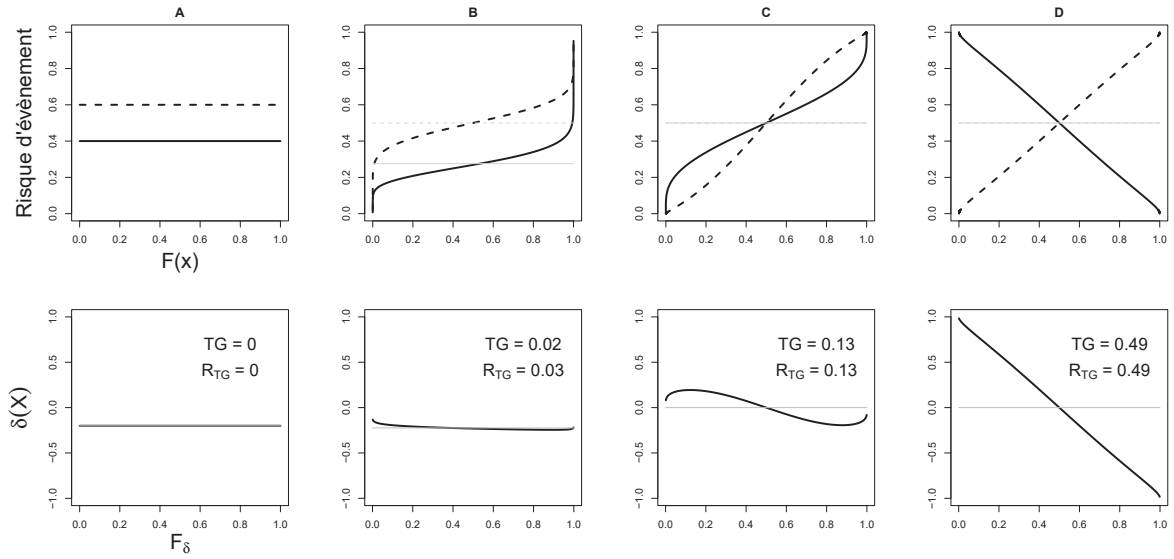


FIGURE 2.9 – Correspondance entre les courbes de risque de quatre marqueurs simulés et le gain total
Courbe pleine : Traitement innovant (graphiques du haut) ; Courbe en pointillés : Traitement de référence (graphiques du haut) ; Courbe pleine noire : $\delta(X)$ (graphiques du bas) ; Courbe pleine grise : $\rho_1 - \rho_0$ (graphiques du bas)

égal à 0.13. Enfin, le marqueur D était un marqueur pour lequel la différence de risques variait grandement avec les valeurs du marqueur, indiquant que celui-ci avait une capacité prédictive potentiellement plus importante que celle du marqueur C. Cela se vérifie avec le TG associé étant égal à 0.49, ce qui est bien supérieur à celui du marqueur C.

Comme souligné par les auteurs, l'interprétation clinique de cet indicateur est difficile. Néanmoins, il a l'avantage d'être borné entre 0 et 1. Si TG $= 0$ alors le marqueur n'a aucune capacité prédictive, si TG $= 1$ alors il s'agit d'un marqueur pour lequel la variabilité de la différence de risques est maximale.

Cependant, les auteurs ne mentionnent pas que la valeur maximale du TG dépend des risques moyens dans chaque bras de traitement. En effet, le TG ne pourra atteindre la valeur de 1 que dans le cas où les risques moyens dans les deux bras sont égaux à 0.5. Dans les autres cas, la valeur maximale du TG dépend des risques moyens de survenue d'évènement dans chaque bras, ce qui complique son interprétation. Il a été déterminé dans le cadre de cette thèse la valeur maximale atteignable par le TG en fonction des risques moyens dans chaque bras.

En effet, en reprenant la définition du marqueur « parfait » donnée dans la sous-section 2.1.2, celui-ci se caractérise par le fait que la quantité q_2 (proportion de patients pour lesquels l'évènement d'intérêt survient avec les deux traitements) ou q_3 (proportion de patients pour lesquels l'évènement d'intérêt ne survient avec aucun des traitements) est nulle. Ces relations permettent d'identifier la valeur maximale du TG, les développements sont présentés en Annexe A, et permettent d'obtenir les résultats suivants :

1. Si $\rho_1 < 1-\rho_0$
   (i) Si $\rho_0 > \rho_1$ alors
   $$\max(\text{TG}) = 2(\rho_0 + \rho_0\rho_1 - \rho_0^2).$$
   (ii) Si $\rho_0 < \rho_1$ alors
   $$\max(\text{TG}) = 2(\rho_1 + \rho_0\rho_1 - \rho_1^2).$$
   (iii) Si $\rho_0 = \rho_1 = \rho$ alors
   $$\max(\text{TG}) = 2\rho.$$

2. Si $\rho_1 > 1-\rho_0$
   (i) Si $\rho_0 > \rho_1$ alors
   $$\max(\text{TG}) = 2(1-\rho_0 + \rho_0\rho_1 - \rho_1^2).$$
   (ii) Si $\rho_0 < \rho_1$ alors
   $$\max(\text{TG}) = 2(1-\rho_1 + \rho_0\rho_1 - \rho_0^2).$$
   (iii) Si $\rho_0 = \rho_1 = \rho$ alors
   $$\max(\text{TG}) = 2(1-\rho).$$

3. Si $\rho_1 = 1-\rho_0$
   $$\max(\text{TG}) = 4(\rho_0 - \rho_0^2).$$

La Figure 2.10 permet de visualiser la valeur de max(TG) en fonction des valeurs de ρ_0 et ρ_1. On constate notamment que max(TG) se rapproche de 0 lorsque les risques moyens dans

chaque bras de traitement s'éloignent de 0.5, ce qui rend l'interprétation du TG difficile.

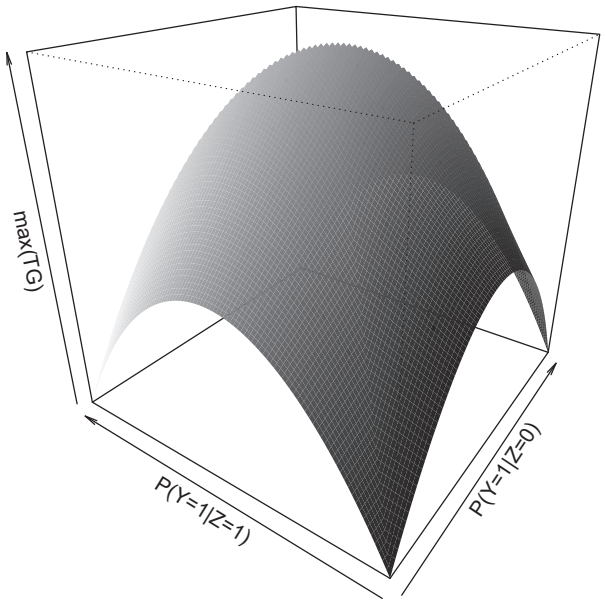


FIGURE 2.10 – Représentation de $\max(\mathrm{TG})$ en fonction des valeurs de ρ_0 et ρ_1

Il pourrait alors être suggéré la construction d'un nouvel indicateur standardisé du type

$$\mathrm{R_{TG}} = \frac{\mathrm{TG}}{\max(\mathrm{TG})},$$

qui permettrait de comparer plusieurs marqueurs en conservant une échelle commune pour quantifier leur caractère prédictif global. Néanmoins, que ce soit le TG ou le $\mathrm{R_{TG}}$, l'estimation de ces indicateurs repose sur des modèles permettant de prédire le risque de survenue de l'évènement d'intérêt dans chacun des bras de traitement. Janes et al. (2014a) proposent d'utiliser un modèle de régression logistique incluant une interaction marqueur-traitement. Il est donc nécessaire que les hypothèses sur lesquelles repose le modèle soient vérifiées pour que l'estimation du TG soit valide.

2.4.2 Courbe d'impact de sélection

La courbe d'impact de sélection d'un traitement (Song and Pepe, 2004) est un outil graphique permettant d'évaluer l'impact qu'aura le marqueur si le choix du traitement repose sur ses valeurs. Contrairement au TG présenté dans la sous-section précédente, la courbe d'impact de sélection repose donc sur une règle d'allocation du traitement. Pour la suite on fera donc l'hypothèse que le traitement innovant ($Z = 1$) est recommandé pour les patients dont la valeur de marqueur est supérieure à un seuil c, et que le traitement de référence ($Z = 0$) est recommandé sinon. Les développements qui suivent peuvent facilement être adaptés si ce n'est pas le cas. La

construction de la courbe d'impact de sélection est basée sur le calcul de l'indicateur suivant :

$$\theta(c) = 1 - \mathrm{P}(Y = 1 | [(X > c) \cap Z = 1] \cup [(X \leq c) \cap Z = 0]) .$$

Il s'agit de la probabilité de ne pas développer l'évènement d'intérêt pour les patients respectant la règle d'allocation définie précédemment. Comme l'étude de cet indicateur est réalisée dans le cadre d'un essai clinique randomisé comparant deux traitements avec une allocation initiale de 1:1, $\theta(c)$ peut s'exprimer comme :

$$\theta(c) = 1 - \left[\mathrm{P}(Y^{(1)} = 1 | X > c)\mathrm{P}(X > c) + \mathrm{P}(Y^{(0)} = 1 | X \leq c)\mathrm{P}(X \leq c) \right] .$$

Plutôt que de représenter l'évolution de l'indicateur $\theta(c)$ en fonction de plusieurs valeurs du seuil c, les auteurs ont choisi de la représenter par rapport à la fonction de répartition du marqueur X évaluée en c. La Figure 2.11 représente deux graphiques. Le premier permet d'interpréter la courbe d'impact de sélection, le second permet de caractériser ce qu'est un marqueur prédictif en utilisant les courbes d'impact de sélection de trois marqueurs différents.

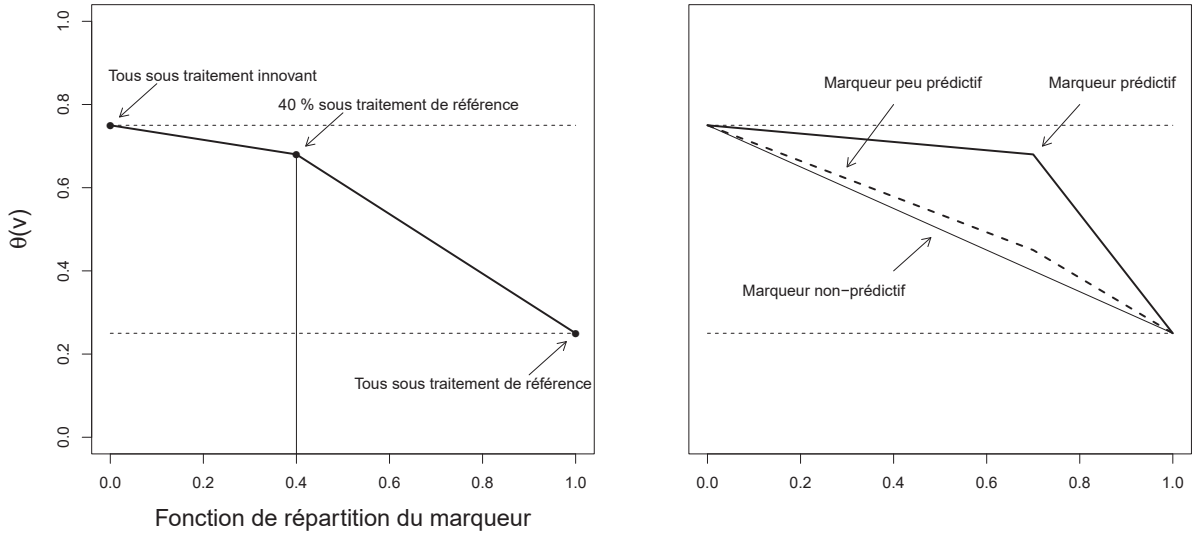


FIGURE 2.11 – Représentation schématique des courbes d'impact de sélection

Bien que le calcul de $\theta(c)$ soit basé sur une règle de décision (et donc sur la définition d'un seuil sur l'échelle du marqueur), la courbe d'impact de sélection permet de représenter l'évolution de $\theta(c)$ pour un ensemble de seuils différents, ce qui permet d'avoir un aperçu global de la capacité prédictive du marqueur. Ainsi, plus la courbe d'impact de sélection d'un marqueur s'éloigne de la « diagonale » tracée sur le graphique, plus sa capacité prédictive est importante. Les auteurs de l'article n'ont pas fait de démonstration de cette affirmation, mais elle peut se faire aisément en réécrivant $\theta(c)$ comme :

$$\theta(c) = 1 - \int_{c}^{+\infty} \rho_1(x)\mathrm{P}(X = x)\, dx - \int_{-\infty}^{c} \rho_0(x)\mathrm{P}(X = x)\, dx$$

$$
\begin{aligned}
&= 1 - \int_{-\infty}^{+\infty} \rho_1(x)\mathrm{P}(X = x)\, dx + \int_{-\infty}^{c} \rho_1(x)\mathrm{P}(X = x)\, dx - \int_{-\infty}^{c} \rho_0(x)\mathrm{P}(X = x)\, dx \\
&= 1 - \rho_1 + \int_{-\infty}^{c} \mathrm{P}(X = x)[\rho_1(x) - \rho_0(x)]\, dx \\
&= 1 - \rho_1 + \int_{-\infty}^{c} \mathrm{P}(X = x) \times \delta(x)\, dx .
\end{aligned}
$$

Lorsque le marqueur n'a pas de capacité prédictive, alors $\delta(X) = \Delta$ et ne dépend plus de X. Il est donc possible d'écrire :

$$
\begin{aligned}
\theta(c) &= 1 - \rho_1 + \int_{-\infty}^{c} \mathrm{P}(X = x) \times \Delta\, dx \\
&= 1 - \rho_1 + (\rho_1 - \rho_0) \times F(c) .
\end{aligned}
$$

Dans ce cas précis, $\theta(c)$ s'exprime comme une fonction affine de $F(c)$ prenant comme valeur $1 - \rho_1$ lorsque $F(c) = 0$ et comme valeur $1 - \rho_0$ lorsque $F(c) = 1$. Ainsi, la diagonale sur les graphiques présentés dans la Figure 2.11 fait bien référence a un marqueur n'ayant pas de capacité prédictive.

La Figure 2.12 permet de visualiser la correspondance entre les courbes de risque présentées initialement dans la Figure 2.1 et les courbes d'impact de sélection.

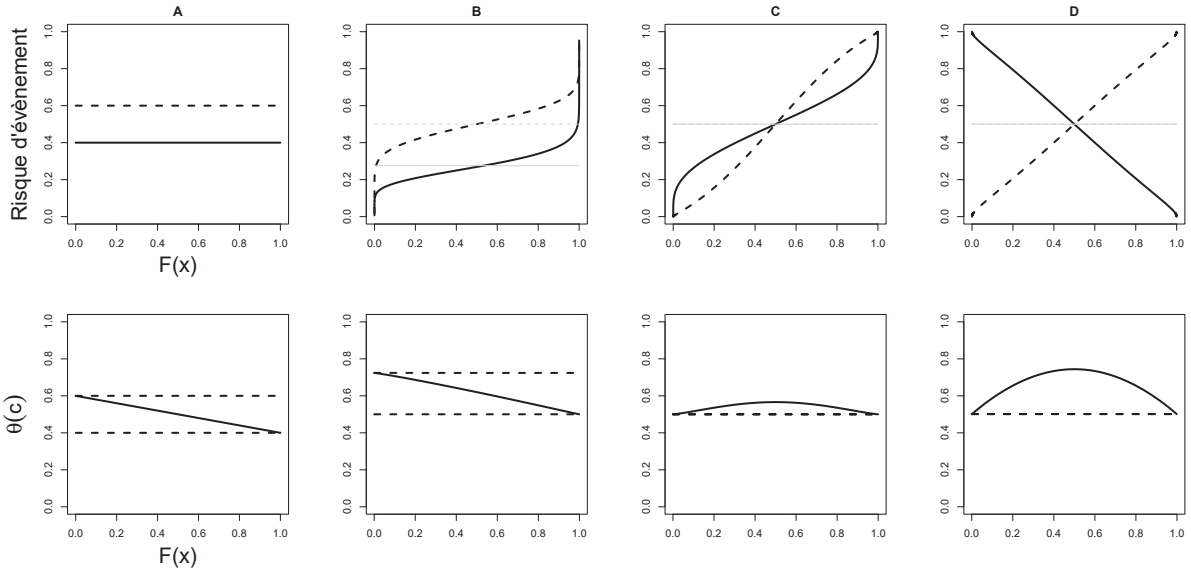


FIGURE 2.12 – Correspondance entre les courbes de risque de quatre marqueurs simulés et les courbes d'impact de sélection
Courbe pleine : Traitement innovant (graphiques du haut) ; Courbe en pointillés : Traitement de référence (graphiques du haut) ; Courbe pleine : courbe d'impact de sélection (graphiques du bas)

On constate que pour les marqueurs A et B qui n'avaient pas (ou très peu) de capacité prédictive, la courbe d'impact de sélection est située sur la diagonale du rectangle formé par les droites en pointillés. Cela indique que ce sont effectivement des marqueurs n'ayant aucune capacité prédictive. Pour le marqueur C, la courbe d'impact de sélection s'éloigne de la droite en

pointillés (les risques moyens dans chaque bras étant égaux à 0.5 il n'y a plus de rectangle, mais seulement une droite) ; cela indique que ce marqueur a une capacité prédictive même si celle-ci semble visuellement assez faible. Enfin, pour le marqueur D, la courbe d'impact de sélection s'éloigne beaucoup de la droite en pointillés ; cela indique que le marqueur a une forte capacité prédictive.

Afin de pouvoir tracer ces courbes, il faut modéliser le risque de survenue de l'évènement observé en fonction des valeurs du marqueur étudié dans chaque bras de traitement en introduisant une interaction marqueur-traitement. Cette approche souffre donc des mêmes limites que le TG et que les approches reposant sur un modèle permettant de prédire le risque de survenue de l'évènement d'intérêt (ex : hypothèse de linéarité des effets du modèle sur l'échelle du prédicteur linéaire). De plus, il s'agit ici d'une approche purement graphique permettant d'avoir un aperçu visuel de la capacité prédictive globale du marqueur, mais pas de quantifier cette capacité globale.

Les auteurs ont ensuite défini plusieurs indicateurs dérivés de cette courbe. Cependant, ce ne sont pas des indicateurs de la capacité prédictive globale du marqueur, car ce sont des mesures évaluant l'impact de l'utilisation du marqueur après la définition d'une règle de décision basée sur un seuil du marqueur. Cette approche a également été adaptée à la prise en compte des critères de jugement du type « temps jusqu'à évènement » par Song and Zhou (2011).

2.4.3 Indice de concordance

Dans leur article, Zhang et al. (2017) proposent une métrique basée sur la valeur du marqueur X et la différence de risques conditionnelle $\delta(X) = \mathbf{E}[Y^{(1)} - Y^{(-1)}|X]$. On notera que les auteurs proposent de coder les bras de traitements en 1 (traitement innovant) et -1 (traitement de référence), contrairement à ce qui est proposé dans d'autres approches, pour des raisons techniques expliquées par la suite.

La mesure de concordance s'exprime comme $\gamma = \mathbf{E}[G_{ij}]$ avec

$$G_{ij} = -\text{sgn}(X_i - X_j)[\delta(X_i) - \delta(X_j)]\,,$$

où X_i et X_j sont les valeurs de marqueur de deux patients indépendants.

Il est possible de constater ici que lorsque $\delta(X) = \Delta$ et ne dépend plus de X, alors $\delta(X_i) = \delta(X_j)$ pour toutes les paires de patients $\{i, j\}$, et $\gamma = 0$.

Ainsi, des valeurs éloignées de 0 sont préférables pour γ. On note également que, puisqu'il s'agit d'une statistique de rang, elle est invariante à toute transformation monotone croissante de X.

Les auteurs ont également pu mettre en évidence un lien intéressant entre la mesure de concordance γ et la courbe d'impact de sélection de Song and Pepe (2004).

Si on fait l'hypothèse, sans perte de généralité de la démonstration, que X est distribué uni-

formément sur l'intervalle $[0;1]$ alors γ peut s'exprimer comme :

$$\begin{aligned}\gamma &= -4 \times \text{Cov}[X, \delta(X)] \\ &= 4\int_0^1 \theta(x)\, dx - 2\left[1 - \text{P}(Y^{(-1)} = 1)\right] - 2\left[1 - \text{P}(Y^{(1)} = 1)\right].\end{aligned}$$

Dans le cas d'un marqueur uniforme, l'indicateur γ peut donc s'interpréter comme un terme de covariance entre X et $\delta(X)$, ce qui explique pourquoi $\gamma = 0$ lorsqu'un marqueur n'a pas de capacité prédictive. En revanche, pour que la réciproque soit vraie, il est nécessaire de faire l'hypothèse que les valeurs de $\delta(X)$ évoluent de manière monotone avec les valeurs de X (hypothèse faite par les auteurs).

Si X n'est pas distribué uniformément, on peut quand même exprimer γ en fonction de la courbe d'impact de sélection sous la forme suivante :

$$\gamma = 4\int_{-\infty}^{+\infty} \theta(x)\text{P}(X = x)\, dx - 2\left[1 - \text{P}(Y^{(-1)} = 1)\right] - 2\left[1 - \text{P}(Y^{(1)} = 1)\right].$$

Comme souligné par les auteurs, γ est fonction de l'aire sous la courbe d'impact de sélection. Cependant, les auteurs ne vont pas plus loin dans l'interprétation de cet indicateur. Pour aller plus loin, si on note l'aire sous la courbe d'impact de sélection d'un marqueur non prédictif sous la forme

$$\mathcal{A}_{NP} = [1 - \text{P}(Y^{(-1)} = 1)]/2 + [1 - \text{P}(Y^{(1)} = 1)]/2,$$

et l'aire sous la courbe d'impact de sélection du marqueur étudié sous la forme

$$\mathcal{A}_X = \int_{-\infty}^{+\infty} \theta(x)\text{P}(X = x)\, dx,$$

il est alors possible d'écrire γ comme :

$$\gamma = 4 \times [\mathcal{A}_X - \mathcal{A}_{NP}].$$

Au regard de cette expression, γ peut s'interpréter (à une constante multiplicative près) comme l'aire entre les courbes d'impact de sélection du marqueur étudié et d'un marqueur non prédictif. Cette interprétation, non relevée par les auteurs dans leur article, permet de mieux comprendre ce que mesure γ.

Estimation de γ

L'estimation de γ est compliquée par le fait que $\delta(X)$ n'est jamais observée puisque chaque patient ne reçoit qu'un seul des deux traitements. Les auteurs proposent alors d'estimer $\delta(X)$ par $\mathbf{E}(2ZY|X)$. On note en effet que

$$\begin{aligned}\mathbf{E}(2ZY|X) &= 2[\mathbf{E}(ZY|X, Z = 1)\text{P}(Z = 1|X) + \mathbf{E}(ZY|X, Z = -1)\text{P}(Z = -1|X)] \\ &= 2[\mathbf{E}(Y|X, Z = 1) \times 0.5 - \mathbf{E}(Y|X, Z = -1) \times 0.5]\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}(Y|X, Z=1) - \mathbf{E}(Y|X, Z=-1) \\
&= \delta(X)\,,
\end{aligned}
$$

où la deuxième ligne est obtenue car la mesure de cette quantité est réalisée dans le cadre d'un essai clinique randomisé avec une allocation de traitement de 1:1 (il est possible de modifier la formule pour d'autres designs d'allocation). En réalité, $\delta(X)$ peut aussi être estimée par $\mathbf{E}[2Z(Y-A)|X]$ pour n'importe quel $A = a(\boldsymbol{W})$ car $\mathbf{E}[\mathbf{E}(ZA|\boldsymbol{W})|X] = 0$ grâce à la randomisation de l'essai clinique. L'introduction de A est justifiée par un gain dans l'efficacité de l'estimateur explicité plus tard.

Ainsi l'estimateur de γ proposé est :

$$\widehat{\gamma}(a) = \frac{2}{n(n-1)} \sum_{i<j} \widehat{G}_{ij}(a)\,,$$

où

$$\widehat{G}_{ij}(a) = -2\mathrm{sgn}(X_i - X_j)[Z_i(Y_i - A_i) - Z_j(Y_j - A_j)]\,.$$

Comme $\widehat{\gamma}(a)$ est une U-statistique d'ordre 2 (Hoeffding, 1948), $\sqrt{n}[\widehat{\gamma}(a) - \gamma(a)]$ converge vers une loi normale centrée sur 0 et de variance

$$\sigma^2(a) = 4 \times \mathrm{Cov}[\widehat{G}_{12}(a), \widehat{G}_{13}(a)]\,.$$

Un estimateur consistent de $\sigma^2(a)$ est donné par :

$$\widehat{\sigma}^2(a) = \frac{8}{n(n-1)(n-2)} \sum_{j \neq i \neq k, j<k} \widehat{G}_{ij}(a)\widehat{G}_{ik}(a) - 4\widehat{\gamma}^2(a)\,.$$

Il est possible de démontrer que la variance asymptotique de l'estimateur $\sigma^2(a)$ est minimisée pour $A = \mathbf{E}(Y|\boldsymbol{W})$. Les auteurs démontrent par la suite que $\widehat{\gamma}(a)$ ne dépend pas du choix de A, en revanche l'estimateur de la variance de $\widehat{\gamma}(a)$ est grandement impacté par ce choix. Le calcul de A repose sur les prédictions d'un modèle (logistique ou probit) dont les hypothèses peuvent alors impacter les performances de la méthode d'inférence de cette métrique. Ainsi, si l'estimateur de γ est non paramétrique, l'optimisation de sa variance peut nécessiter le recours à un modèle paramétrique. Par ailleurs, une autre limite de cet indicateur est qu'il n'a pas d'échelle universelle (car son calcul repose sur les risques moyens dans chaque bras de traitement), ce qui limite les comparaisons entre marqueurs de différentes études.

La Figure 2.13 permet de visualiser la correspondance entre les courbes de risque présentées initialement dans la Figure 2.1 et l'indice de concordance γ.

On constate que pour le marqueur A qui n'avait pas de capacité prédictive, alors $\gamma = 0$ confirmant ainsi que ce marqueur n'est pas utile pour guider le choix du traitement. Pour le marqueur B, la courbe d'impact de sélection était très proche de la diagonale, indiquant le peu d'intérêt de ce marqueur, cela se confirme avec un $\gamma = 0.02$ ici. Pour le marqueur C, $\gamma = 0.16$ indiquant que ce marqueur a une capacité prédictive non nulle, ce qui est en adéquation avec le résultat

des autres méthodes. Enfin, concernant le marqueur D, il a une valeur de $\gamma = 0.65$ bien supérieure à celle du marqueur C confirmant ainsi que le marqueur D a une capacité prédictive plus importante que le marqueur C.

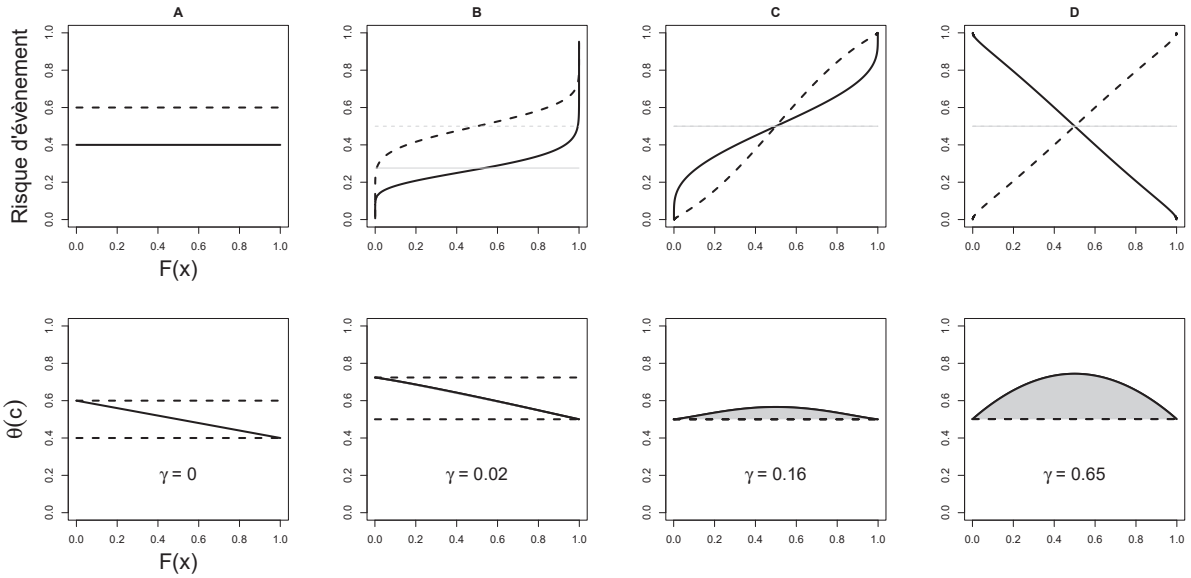


FIGURE 2.13 – Correspondance entre les courbes de risque de quatre marqueurs simulés et l'indice de concordance γ
Courbe pleine : Traitement innovant (graphiques du haut) ; Courbe en pointillés : Traitement de référence (graphiques du haut) ; Courbe pleine : courbe d'impact de sélection (graphiques du bas) ; Aire grisée : $\gamma/4$

2.5 Proposition d'un nouvel indicateur : l'ABC

Exception faite de l'indicateur proposé par Zhang et al. (2017), les indicateurs permettant de mesurer la capacité prédictive globale d'un marqueur reposent pour la plupart sur la modélisation du risque de survenue de l'évènement étudié dans chacun des bras de traitement. D'autres auteurs proposent également de modéliser directement $\delta(X) = \mathbf{E}[Y^{(1)} - Y^{(0)}|X]$, (Huang et al., 2012; Zhang et al., 2014) cependant nous avons pu voir que cette modélisation reposait sur des hypothèses parfois peu réalistes, et dans tous les cas très difficiles à vérifier dans la pratique (Janes et al., 2015b).

Toutes ces remarques ont motivé le développement d'un indicateur ne reposant pas sur la modélisation du risque de survenue de l'évènement étudié dans le cas particulier où aucun traitement n'est meilleur que l'autre en moyenne (c'est-à-dire que $\rho_0 = \rho_1$). En effet, c'est généralement dans ce genre de situation que la recherche d'un marqueur prédictif est principalement réalisée.

Il a été présenté précédemment que la recherche d'un marqueur prédictif est synonyme de la recherche d'une interaction entre le marqueur et les traitements étudiés sur l'échelle additive du risque de survenue d'évènement. Une telle définition revient donc à rechercher un différentiel d'effet pronostique du marqueur entre chacun des bras de traitement.

Une approche classique afin de mesurer le caractère pronostique d'un marqueur pour un bras de traitement donné est de tracer la courbe ROC associée et de calculer l'aire sous la courbe ROC (Hanley and McNeil, 1982). Un moyen de mesurer le différentiel d'effet pronostique du marqueur entre les deux bras de traitement peut donc être de calculer l'aire entre les courbes ROC de chaque traitement (appelée ABC ci-après). Lorsque les courbes ROC ne se croisent pas l'ABC peut être mesurée comme étant la différence entre les AUCs des courbes ROC de chaque bras de traitement :

$$\Delta_\theta = \theta_1 - \theta_0 \,,$$

où θ_1 et θ_0 sont les AUCs des courbes ROC pour le bras innovant et de référence, respectivement. Lorsqu'elles se croisent, des méthodes permettant d'estimer des AUCs partielles (McClish, 1989) pourraient être utilisées, mais cela nécessiterait des développements supplémentaires.

La Figure 2.14 permet de comparer les courbes de risque issues de quatre marqueurs différents avec les courbes ROC correspondantes. Ces quatre marqueurs simulés ont été construits de telle sorte que les risques moyens dans chaque bras sont égaux.

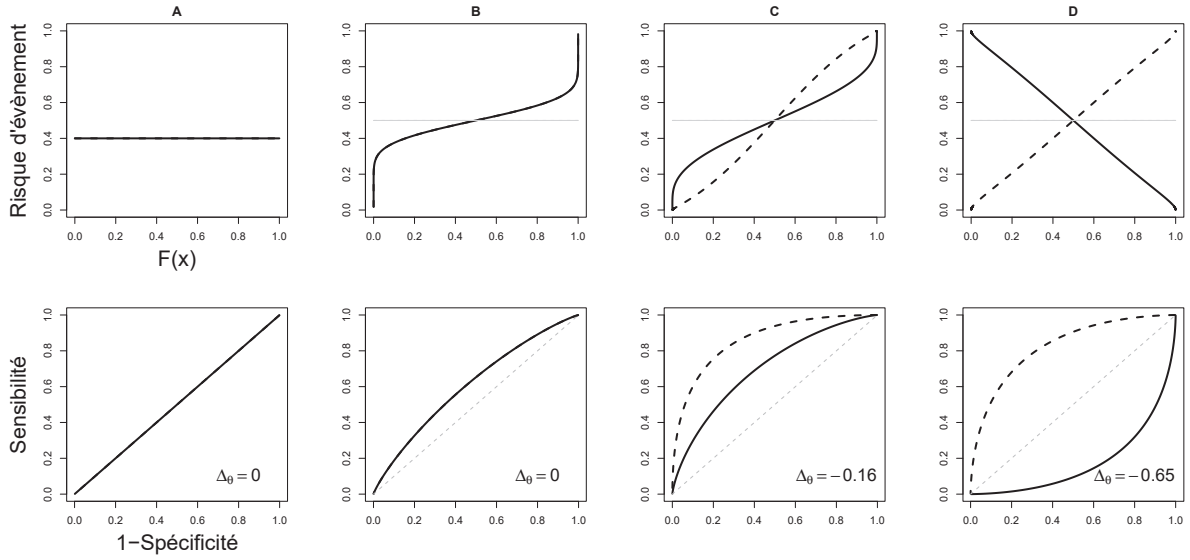


FIGURE 2.14 – Correspondance entre les courbes de risque pour quatre marqueurs simulés et les courbes ROC dans chaque bras de traitement
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

Pour le marqueur A, les courbes de risque dans chaque bras sont horizontales et superposées. Ce marqueur n'a aucune capacité prédictive, et cela se vérifie avec les courbes ROC associées qui sont toutes deux superposées et sur la diagonale. Dans ce cas $\Delta_\theta = 0$ et le marqueur n'a effectivement aucune capacité prédictive. Pour le marqueur B, les courbes de risque dans chaque bras évoluent en fonction des valeurs du marqueur et sont superposées. Un tel marqueur ne doit avoir aucune capacité prédictive puisque la différence de risque est nulle et constante en fonction des valeurs du marqueur. Cela se vérifie avec les courbes ROC associées qui sont superposées. Dans ce cas $\Delta_\theta = 0$ et le marqueur n'a aucune capacité prédictive. Pour le marqueur C, $\delta(X)$

varie en fonction des valeurs du marqueur, il s'agit donc à première vue d'un marqueur ayant une capacité prédictive quantifiée par $\Delta_\theta = -0.16 \neq 0$. Ce marqueur a donc bien une capacité prédictive. Enfin le marqueur D est un marqueur pour lequel les variations dans la fonction $\delta(X)$ sont plus importantes que pour le marqueur C. Cela se reflète dans la mesure $\Delta_\theta = -0.65$ qui est plus éloignée de 0 que l'aire entre les courbes ROC du marqueur C. Le marqueur D a lui aussi une capacité prédictive.

2.5.1 Article soumis dans la revue Biometrical Journal

Cet article présente de manière approfondie cet indicateur et la méthode d'estimation proposée, détaille sa justification théorique, évalue et compare ses performances à celles des autres indicateurs évoqués dans ce chapitre. Les informations supplémentaire soumises à cette revue sont disponibles en Annexe B.

The area between ROC curves, a non-parametric method to evaluate a biomarker for patient treatment selection

Yoann Blangero* [1,2], **Muriel Rabilloud** [1,2], **Pierre Laurent-Puig** [3,4,5], **Karine Le Malicot** [6], **Côme Lepage** [6,7,8], **René Ecochard** [1,2], **Julien Taieb** [3,9], and **Fabien Subtil** [1,2]

[1] Service de Biostatistique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France

[2] Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Villeurbanne, France

[3] Université Paris Descartes, Sorbonne Paris Cité, Paris, France

[4] Service de génétique, Hôpital Européen Georges Pompidou, Paris, France

[5] INSERM UMR-S 1147, Paris, France

[6] Fédération Francophone de Cancérologie Digestive, Dijon, France

[7] Hépato-gastroentérologie et cancérologie digestive, Centre hospitalier universitaire Dijon Bourgogne, Dijon, France

[8] INSERM U 866, Dijon, France

[9] Chirurgie digestive générale et cancérologique, Hôpital Européen Georges Pompidou, Paris, France

Treatment selection markers are generally sought for when the benefit of an innovative treatment in comparison with a reference treatment is considered, and that this benefit is suspected to vary according to the characteristics of the patients. Classically, such quantitative markers are detected through testing a marker-by-treatment interaction in a parametric regression model. Most alternative methods rely on modelling the risk of event occurrence in each treatment arm or the benefit of the innovative treatment over the marker values, but with assumptions that may be difficult to verify. Herein, a simple non-parametric approach is proposed to detect and assess the general capacity of a quantitative marker for treatment selection when no overall difference in efficacy could be demonstrated between two treatments in a clinical trial. This graphical method relies on the area between treatment-arm-specific ROC curves (ABC), which reflects the treatment selection capacity of the marker. A simulation study assessed the inference properties of the ABC estimator and compared them with other parametric and non-parametric indicators. The simulations showed that the estimate of the ABC had low bias, power comparable to parametric indicators, and that its confidence interval had a good coverage probability (better than the other non-parametric indicator in some cases). Thus, the ABC is a good alternative to parametric indicators. The ABC method was applied to data of the PETACC-8 trial that investigated FOLFOX4 vs. FOLFOX4 + cetuximab in stage III colon adenocarcinoma. It enabled the detection of a treatment selection marker: the *DDR2* gene.

Key words: clinical trial; predictive marker; quantitative marker; receiver operating characteristic curve; treatment selection

1 Introduction

A major aim of precision medicine is to determine the best treatment for individual patients. It is therefore essential to identify and assess markers able to guide treatment decisions so as to avoid the occurrence of a given event (e.g. disease progression, recurrence, or death) in a given post-treatment interval. When comparing the efficacy of two treatments (an innovative vs. a reference treatment), such markers are expected to improve patient outcomes by selecting patients who would likely most benefit from the innovative treatment and avoid treating those who would not benefit from this. There is currently no consensus on

*Corresponding author: e-mail: yoann.blangero@chu-lyon.fr, Phone: +33472115751, Fax: +33472115141

the naming of such a marker; whereas Italiano (2011) and Ballman (2015) have used "predictive marker" Janes et al. (2014a) used "treatment selection marker" that will be used herein.

A treatment selection marker is generally sought for when the overall risk of event occurrence is nearly the same with two different treatments; it is then expected that a subgroup of patients would get more benefit from one of the treatment than from the other. One example of treatment selection marker is the mutated KRAS gene in metastatic colorectal cancer. The presence of this mutated gene is a marker of benefit from chemotherapy alone as opposed to chemotherapy + epidermal growth factor receptor –EGFR– inhibitor; patients with tumors harboring mutated KRAS exon 2 are known to be resistant to EGFR inhibitors, whereas those with KRAS wild-type tumors do benefit from the combined treatment (Di Fiore et al., 2007; Lièvre et al., 2008; De Roock et al., 2008).

In the case of a quantitative marker, it is necessary to find a threshold value of the marker that determines the optimal treatment allocation for the patients with a marker value above or below this threshold (Vickers et al., 2007; Janes et al., 2014a,b; Blangero et al., 2019). However, before defining a threshold, the first step in the assessment of a new promising quantitative marker is to quantify and test its overall performance for treatment selection. Various methods have been proposed to evaluate the overall performance of a marker for treatment selection. The classical approach consists in modelling the risk of event given the treatment options and the marker values, and then testing for a statistical interaction between these two variables, as proposed by Byar (1985), and applied in several studies (for some examples, see Weidhaas et al. (2016), or Skougaard et al. (2016)). One limit of this approach is that the interaction coefficient depends on the additive or multiplicative structure of the model; the interaction may be present in one type of model but not in the other, and conversely so (Byar, 1985). A marker is defined as a treatment selection marker when the difference in risk of event occurrence between the two treatment arms is inconstant over the marker values (Song and Pepe, 2004), which means that the additive scale should be used to assess treatment selection markers.

In addition, although the interaction approach is straightforward with binary markers, it is quite complex with quantitative markers because of the difficulty of verifying the adequacy of the functional form retained in modeling the interaction. One extension of the previous approach is the use of graphical tools such as "marker-by-treatment predictiveness curves" as proposed by Janes et al. (2011). These graphs plot the risk of event in each treatment arm given the marker value vs. the cumulative distribution function of the marker. The cumulative distribution function instead of real values enables the use of a single scale ranging between 0 and 1, allows marker-by-treatment predictiveness curve comparisons, and gives the proportions of patients who would receive each treatment according to the marker values, which is important in medical decision-making. Marker-by-treatment predictiveness curves allow visualization of the performance of a marker for treatment selection, but they rely on a good calibration of the risk modeling in each arm. Such a model often assumes a linear marker-by-treatment interaction on the linear predictor scale. Unfortunately, this assumption is not always valid and not easy to check. Moreover, the marker-by-treatment predictiveness curves is a graphical tool, but does not allow to quantify the performance of the marker for treatment selection.

Other methods assess the treatment selection capacity of a quantitative marker (Huang et al., 2012; Zhang et al., 2014) by measuring its ability to distinguish patients who would have a better outcome with the innovative treatment in comparison to the reference treatment, from patients who would have a worse outcome. However, this kind of approach needs to model the probability of having a better outcome with the innovative treatment compared with the reference treatment in each patient. Except in cross-over trials, this requires modeling using a potential outcomes framework with complex assumptions that may be very difficult to verify (Janes et al., 2015b). For example, Huang et al. (2012) made the monotonicity assumption (one treatment is always at least as effective as the other one) to estimate the individual benefit, which is a strong assumption. Zhang et al. (2014) relaxed the latter assumption by assuming that the potential outcomes are independent given observed covariates. That means that the benefit of the innovative treatment for a patient may be calculated by comparing its outcome to the one of patients similar regarding the

observed covariates but receiving the reference treatment. This method assumes that the observed covariates are sufficient to explain the dependence between the two potential outcomes, which is an assumption difficult to verify.

In the present paper, a simple non-parametric method is proposed to investigate the capacity of a quantitative marker for treatment selection when the overall risk of event in each treatment arm is equal. The method relies on a special use of Receiver Operating Characteristic (ROC) curves and provides a bounded indicator able to quantify and test the treatment selection capacity of the marker. The method is described, tested, and compared with other methods in a simulation study; it is then applied to a real dataset.

2 Methods

Throughout this article it is assumed that the marker under study (denoted V) is measured before treatment allocation within the context of a parallel randomized controlled clinical trial with two treatment arms (innovative vs. reference) and that the outcome of interest is a binary event measured after a fixed duration of follow-up.

The binary event of interest is denoted by E, where $E = 1$ indicates the presence of the event of interest, and $E = 0$ its absence. Let us also denote the treatments under study by T, where $T = 1$ indicates the innovative treatment and $T = -1$ indicates the reference one.

Moreover, it is assumed that:

$$\rho_{(-1)} = \rho_{(1)} = \rho \tag{1}$$

where $\rho_{(-1)} = \mathrm{P}(E = 1|T = -1)$ and $\rho_{(1)} = \mathrm{P}(E = 1|T = 1)$ denote the overall risk of event in each arm, and $\rho = \mathrm{P}(E = 1)$ denotes the marginal risk of event in the trial. Assumption (1) means that no overall difference in efficacy could be demonstrated between the two treatment arms.

Song and Pepe (2004) proposed a mathematical definition of a treatment selection marker. A marker has no capacity for treatment selection if

$$\delta(v) = \rho_{(-1)}(v) - \rho_{(1)}(v) = \rho_{(-1)} - \rho_{(1)} \qquad \forall v \tag{2}$$

where $\rho_{(-1)}(v) = \mathrm{P}(E = 1|T = -1, V = v)$ and $\rho_{(1)}(v) = \mathrm{P}(E = 1|T = 1, V = v)$ denote the risk of event in each treatment arm for a value v of the marker.

Conversely, a marker has a capacity for treatment selection when the difference in risks between the two treatment arms is dependent of the marker values. As Song and Pepe (2004) and Janes et al. (2014a) suggested, the difference in risk is the key point in treatment selection marker assessment. A marker is all the more interesting for treatment selection that the changes in risk differences are important according to the marker values.

2.1 Marker-by-treatment predictiveness curves

Marker-by-treatment predictiveness curves are simple graphical tools that help understanding the difference between a treatment selection marker and a simple prognostic marker.

In Figure 1, each panel presents two curves: one relative to the reference treatment and another relative to the innovative treatment.

- Panel A shows a case where the risk of event is independent of the marker value in each treatment arm. As the overall risk of event is the same in each treatment arm (assumption (1)), the marker-by-treatment predictiveness curves overlap; hence, the marker cannot be a treatment selection marker.

- Panel B shows a case where the risk changes with the marker value in both treatment arms. Thus, the marker may be called “prognostic marker” in each arm. However, the difference in risk between the two arms ($\delta(V)$) is constant and equal to 0 in this case. Thus, there is no interaction between the treatment arm and the marker values, the marker cannot be a treatment selection marker.

– Panel C shows a case where the risk of event occurrence decreases with the marker value in the innovative arm but increases in the reference arm: the prognostic value is different between treatment arms. $\delta(V)$ changes with the marker value: this marker is thus a treatment selection marker. In this case, the threshold of marker value that defines treatment allocation should be close to the marker value that corresponds to 50% of its cumulative distribution.

– Panel D shows another case where the risk of event occurrence decreases with the marker value in the innovative arm and increases in the reference arm, but the slopes are greater than in panel C: the prognostic value of the marker is stronger in the two treatment arms. This marker is also a treatment selection marker; furthermore, its capacity for treatment selection is greater than in panel C because of greater magnitude of changes in $\delta(V)$. The treatment selection capacity of a marker is all the more important that the changes in $\delta(V)$ over the cumulative distribution function of the marker values are important.

Thus, a marker is a treatment selection marker when its prognostic ability is different between two treatment arms, which is the definition of a marker-by-treatment interaction. This is the basis for the development of the method presented hereafter.

2.2 Notations and illustration of the "area between curves"

A simple and non-parametric method to estimate the prognostic ability of a marker in a single treatment arm relies on the area under the ROC curve (AUC, θ in equations) that quantifies the ability of the marker to discriminate subjects who will experience the event in a given post-treatment interval from those who will not (Hanley and McNeil, 1982). We propose to estimate the treatment selection capacity of a marker by estimating the difference in prognostic ability between two treatment arms. This difference can be quantified by the area that separates the two treatment-arm-specific ROC curves, named "area between curves" (ABC).

A classical assumption in marker evaluation using ROC curves is that the risk of event in both arms is either monotonically increasing or monotonically decreasing over the marker values. Otherwise, the issue of improper ROC curves would arise (Metz and Pan, 1999).

When the two ROC curves do not intersect, the ABC can be measured by the difference between the two AUCs: $\Delta_\theta = \theta_{(-1)} - \theta_{(1)}$, $\theta_{(-1)}$ and $\theta_{(1)}$ being, respectively, the AUCs of the marker for $T = -1$ and $T = 1$. As both $\theta_{(-1)}$ and $\theta_{(1)}$ range between 0 and 1, Δ_θ ranges between -1 and 1 (when Δ_θ is negative, the ABC is the absolute value of Δ_θ).

The second row in Figure 1 presents the ROC curves that correspond to the marker-by-treatment predictiveness curves of the first row.

– Panel E shows two overlapping ROC curves on the diagonal; the marker has no prognostic ability in either arm, $\Delta_\theta = 0$.

– Panel F shows two overlapping ROC curves but distinct from the diagonal; the marker has the same prognostic ability in both arms but no capacity for treatment selection and $\Delta_\theta = 0$.

– Panel G shows two distinct ROC curves located on either side of the diagonal; the marker has a prognostic ability in both arms but the risk is increasing in the innovative arm and decreasing in the reference arm. The marker in Panel G has a capacity for treatment selection and $\Delta_\theta = -0.11$.

– Panel H shows two distinct ROC curves located on either side of the diagonal too. As shown by the marker-by-treatment predictiveness curves, the marker in panel H has a stronger capacity for treatment selection than the marker in panel G. This is reflected by a $\Delta_\theta = -0.48$ further from zero than in panel G.

To summarize, the capacity of a marker for treatment selection increases with the ABC or the gap between the ROC curves; thus Δ_θ different from 0. When $\Delta_\theta = 0$ (overlapping ROC curves) the marker has no capacity for treatment selection. When Δ_θ is equal to -1 or 1, the marker is a perfect treatment selection marker; i.e. the marker distinguishes perfectly patients with (or, alternatively, without) the event under the innovative treatment from those under the reference treatment (Appendix A.1 presents an illustration of the definition of a perfect marker). Furthermore, Δ_θ may be used to test whether a marker has a statistically significant treatment selection capacity (i.e. testing whether $\Delta_\theta = 0$) and compare the capacity for treatment selection of several markers.

2.3 Justification of the use of Δ_θ

The use of Δ_θ to quantify the treatment selection capacity of a marker is justified by its close connection with the difference in risk between the two treatment arms over the marker values. Viallon and Latouche (2011) demonstrated that the AUC in a single treatment arm could be written as a function of the predictiveness curve:

$$\theta_{(T)} = \frac{\int_{-\infty}^{+\infty} F(v)\rho_{(T)}(v)\, dF(v) - \frac{\rho_{(T)}^2}{2}}{\rho_{(T)}(1-\rho_{(T)})}$$

where $F(.)$ is the cumulative distribution function of marker V. With this expression, it is easy to show that when the overall risks of event occurrence in the reference and innovative treatment arms are equal (i.e. when $\rho_{(-1)} = \rho_{(1)} = \rho$), Δ_θ can be expressed as a function of $\delta(V)$:

$$\Delta_\theta = \theta_{(-1)} - \theta_{(1)} = \frac{\int_{-\infty}^{+\infty} F(v) \times \delta(v)\, dF(v)}{\rho(1-\rho)}$$

Δ_θ is greater when the variations in the risk difference $\delta(V)$ are high on the range of marker values, hence when the marker has a greater capacity for treatment selection. According to equation (2), it can be shown that when a marker has no capacity for treatment selection then $\Delta_\theta = 0$, and conversely (Appendix A.2).

2.4 Connection between Δ_θ and two other indicators

Hereafter, two indicators are presented in order to show their connection with Δ_θ: the total gain indicator of Janes et al. (2014b) and the γ indicator of Zhang et al. (2017).

2.4.1 The Total Gain

In their article, Janes et al. (2014a) proposed to evaluate the overall capacity of a marker for treatment selection using the total gain (TG) expressed as:

$$\text{TG} = \int |\delta(v) - (\rho_{(-1)} - \rho_{(1)})|\, dF_\delta$$

In this equation, F_δ is the cumulative distribution function of $\delta(V)$.

The TG indicator measures the overall treatment selection capacity of a marker. When the marker has no treatment selection capacity, the TG equals 0, and conversely so. However, the maximum TG value depends on $\rho_{(-1)}$ and $\rho_{(1)}$, and therefore the TG cannot be used to compare markers from different studies.

The TG and Δ_θ are two closely connected overall indicators of the treatment selection capacity of a marker. From the expressions of TG and Δ_θ, one may see that there is a monotone relationship between these two indicators and that the intensity of this relationship depends on the overall risk of event occurrence in each treatment arm. However, whereas the TG is based on risks, Δ_θ is based on ROC curves that measure the ability of the marker to separate two groups of patients. In fact, Δ_θ is an indicator of the ability

of the marker to distinguish patients with (or, alternatively, without) the event under the innovative treatment from those under the reference treatment. Moreover, Δ_θ is non-parametric regarding the functional form of the interaction, and is always bounded between -1 and 1.

Finally, note that Janes et al. (2014a) did not propose an inference method for the TG except from bootstrap.

2.4.2 The γ concordance measure

In another article, Zhang et al. (2017) proposed a quantitative concordance measure for the assessment of the overall performance of treatment selection markers. This concordance measure is expressed as

$$\gamma = \mathrm{E}(G_{ij})$$

where $G_{ij} = \mathrm{sgn}(V_i - V_j)[\delta(V_i) - \delta(V_j)]$, i and j are the indices of two independent patients, and $\mathrm{sgn}(.)$ is the sign function.

As one may see, when $\delta(V)$ is constant over the marker values then $\gamma = 0$, and the greater the variations in $\delta(V)$ are, the greater γ is, and the greater the performance of the marker for treatment selection is.

There is a connection between this indicator and the two ones described above as there are all functions of $\delta(V)$ and that their value depends on the variations in $\delta(V)$. γ is estimated non-parametrically using pairwise comparisons of patient outcomes; since it is a U-statistic, the estimator converges to a normal distribution (Hoeffding, 1948). The variance of the estimator follows from asymptotic theory and may rely on a working model that predicts the risk of event occurrence in order to be more efficient. The variance is optimal when the working model for risk prediction includes all the covariates that impact the risk of event (see Zhang et al. (2017) for more details).

2.5 Estimation and inference of Δ_θ

When estimated non-parametrically with the trapezoidal rule, the AUC estimate is asymptotically normally distributed with Delong's variance (DeLong et al., 2011). As Δ_θ is a difference between two independent AUCs, its estimator is also asymptotically normally distributed, with variance equal to the sum of the two AUC variances. Thus, a symmetric confidence interval can be obtained using the normal approximation:

$$\left[\widehat{\Delta}_\theta \pm z_{1-\alpha/2} \times \sqrt{\mathrm{Var}(\widehat{\Delta}_\theta)}\right]$$

In this expression $\widehat{\Delta}_\theta$ denotes the estimator of Δ_θ, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

The symmetric confidence interval may indicate limits > 1 or < -1, especially when Δ_θ is close to 1 or -1. To obtain asymmetric confidence limits between 1 and -1, these limits may be calculated on the inverse hyperbolic tangent scale of Δ_θ using the Delta method:

$$\left[\mathrm{arctanh}(\widehat{\Delta}_\theta) \pm z_{1-\alpha/2} \times \frac{\sqrt{\mathrm{Var}(\widehat{\Delta}_\theta)}}{(1+\widehat{\Delta}_\theta)(1-\widehat{\Delta}_\theta)}\right]$$

The confidence interval on the inverse hyperbolic tangent scale of Δ_θ is then back-transformed to provide the asymmetric interval for Δ_θ.

The treatment selection capacity of a marker may be tested using the Wald statistic:

$$z = \frac{\widehat{\Delta}_\theta}{\sqrt{\mathrm{Var}(\widehat{\Delta}_\theta)}}$$

In this formula, z follows asymptotically a standardized normal distribution when the marker has no capacity for treatment selection ($\Delta_\theta = 0$).

The above estimation method is investigated hereafter in a simulation study and applied later to a real dataset.

3 Simulation study

3.1 Design

Simulation studies were designed to evaluate the performance of $\widehat{\Delta}_\theta$ to evaluate the overall capacity of a marker for treatment selection.

To evaluate the method performance, three scenarios for the risk of event occurrence were created. Scenarios 1, 2, and 3 considered that the overall risk in both $T_{(-1)}$ and $T_{(1)}$ was equal to 0.5, 0.25, and 0.1, respectively. Varying the overall risks is expected to affect the variance of Δ_θ because the number of events in each arm changes according to each arm-specific risk. Each scenario was evaluated with four theoretical values of Δ_θ (0.6, 0.4, 0.2, and 0.1; except for Scenario 3 where value 0.6 could not be considered), and five N sizes of 200, 400, 1000, 1600, and 2000 subjects. This generated 55 different settings. Each was run 10 000 times.

As the marker is assessed in a clinical trial context, its values are supposed to have the same distribution in the two trial arms. This is defined as the randomization constraint that can be expressed as:

$$\mathrm{P}(V \leq c|T = -1) = \mathrm{P}(V \leq c|T = 1) \qquad \forall c \tag{3}$$

In Scenario 1, the marker values were sampled from four Gaussian distributions with same variance; two distributions per arm, one for those who will experience the event and one for the others. The means and variances of these distributions were chosen so as to obtain the four theoretical Δ_θ values whilst fulfilling the randomization constraint given in equation (3) (for details, see Supplemental Material). The theoretical Δ_θ values were calculated using the analytical expression of AUC with Gaussian distributions (Pepe, 2003).

The randomization constraint (3) could not be fulfilled with four Gaussian distributions in Scenarios 2 and 3 (see Supplemental Material). Hence, in these scenarios, the marker values were sampled from Gaussian distributions with same variance, except for patients without event in the innovative treatment arm; for these, the marker distribution was built from the distributions in the three other groups using the randomization constraint. A Metropolis algorithm (Gelman et al., 2013) was used to sample values from this compound marker distribution (see Supplemental Material). The means of the three Gaussian distributions were chosen so as to obtain the four theoretical Δ_θ values. The theoretical Δ_θ values were calculated by numerical integration.

A first simulation study was performed to assess the coverage probability of the symmetric and asymmetric confidence intervals of Δ_θ for all Scenarios, Δ_θ values, and N sizes defined previously. Additional simulations were performed with Scenario 1 and a theoretical $\Delta_\theta = 0.95$ to assess the coverage probability of the symmetric and asymmetric confidence intervals when Δ_θ is close to 1.

A second simulation study was performed to estimate the mean relative bias in $\widehat{\Delta}_\theta$, and its power for detection of treatment selection markers. The mean relative bias, the power, and the coverage probability of the asymmetric confidence interval for $\widehat{\Delta}_\theta$ were compared with those of the TG estimator (relative bias, and coverage probability of the percentile bootstrap confidence interval), denoted $\widehat{\mathrm{TG}}$, the γ estimator (relative bias, coverage probability of confidence intervals, and power), denoted $\widehat{\gamma}$, and the interaction coefficient estimator of a logistic regression model (power), denoted $\widehat{\beta}_3$.

For $\widehat{\mathrm{TG}}$, $\delta(V)$ was estimated using the predictions from a logistic regression model assuming a linear interaction between the marker value and the treatments:

$$\mathrm{logit}[\mathrm{P}(E = 1|T, V)] = \beta_0 + \beta_1 \times V + \beta_2 \times T + \beta_3 \times V \times T$$

The power of the interaction coefficient method was calculated using the maximum likelihood estimator of the coefficient β_3 of this model. For the γ concordance measure, the estimation method provided by Zhang et al. (2017) was used based on the predictions of the following working model:

$$\text{logit}[\text{P}(E = 1|V)] = \eta_0 + \eta_1 \times V$$

This working model should lead to an optimal variance estimate as the occurrence of the event of interest only depends on the treatments (that should not be included in this working model) and on the marker under study.

The theoretical TGs and γs were calculated by numerical integration in order to estimate the relative bias in $\widehat{\text{TG}}$ and in $\widehat{\gamma}$.

A third simulation was performed to verify the properties of the Wald test applied to $\widehat{\Delta}_\theta$ and to $\widehat{\beta}_3$ by checking the α-risk value. This simulation was performed with Scenario 1, $\Delta_\theta = 0$, and on five N sizes of 40, 100, 200, 400, and 1000 subjects.

A fourth simulation was performed to assess the impact of deviations from assumption (1) ($\rho_{(-1)} = \rho_{(1)}$) on the estimation and inference properties of $\widehat{\Delta}_\theta$.

3.2 Results

For the first simulation study, in Scenario 1 when $\Delta_\theta = 0.95$, the symmetric confidence interval did not provide a good coverage probability; up to N = 400 inclusive, the mean coverage probability was <93.5%. With larger N (1000, 1600, and 2000), the coverage probability of the symmetric confidence interval was close to 95%. The asymmetric confidence interval provided a better coverage probability; i.e. nearly 95% irrespective of N. Moreover, the mean width of the asymmetric confidence interval was equal to that of the symmetric confidence interval irrespective of N, which means that with the asymmetric confidence interval the gain in coverage probability was not associated with a loss of precision (Table 1).

In other settings, the coverage probabilities of symmetric and asymmetric confidence intervals of Δ_θ were always close to 95% and their mean widths were close in almost all settings. In Scenario 3, the symmetric confidence interval gave coverage probabilities <94% with N = 200 (Tables 1, 2, and 3).

For the second simulation study, the mean relative bias in $\widehat{\Delta}_\theta$ was $< 3 \times 10^{-2}$ in all cases. The relative bias values did not meaningfully change with the risk scenarios (between -0.00417 and 0.03636) but decreased when N increased. The mean relative bias in $\widehat{\text{TG}}$ was also always close to 0; however, in most cases it was higher than the mean relative bias in $\widehat{\Delta}_\theta$, especially when the true Δ_θ and N were equal to 0.1 and 200, respectively. The mean relative bias in $\widehat{\gamma}$ was always close to 0, and close to the mean relative bias in $\widehat{\Delta}_\theta$ (Tables 4, 5, and 6).

The coverage probability of γ was close to 95% in all settings of Scenario 1 (between 93.67% and 95.07%). For Scenario 2, it was close to 95% except for $\Delta_\theta = 0.6$ for which coverage probability was $> 97.5\%$, and for $\Delta_\theta = 0.4$ for which coverage probability was $> 96\%$ irrespective of the sample size. In Scenario 3, with $\Delta_\theta = 0.4$, the coverage probability was always $> 98\%$, while it was close to 95% in other settings (Tables 4, 5, and 6).

The coverage probability of the TG was always close to 95% in all settings of Scenario 1 except when $\Delta_\theta = 0.1$, and $N \leq 400$ ($> 96.9\%$). In Scenario 2, it was close to 95% in all settings, except when $\Delta_\theta = 0.6$ for which the coverage probability of the bootstrap confidence interval decreased with the increase of N (92.86% with $N = 200$, and 64.45% with $N = 2000$). Note that for the latter setting, the TG estimator had a larger bias than in most of the other settings. Also, when $\Delta_\theta = 0.1$ the coverage probability of the bootstrap confidence interval was $> 96.7\%$ with $N \leq 400$. In Scenario 3, it was close to 95% in all settings, except when $\Delta_\theta = 0.1$ and $N \leq 1600$ ($> 96.7\%$).

The power of $\widehat{\Delta}_\theta$ decreased along with the decrease of N and as Δ_θ approached 0. This power varied between risk scenarios but decreased along with the decrease of the risk in each arm. With N = 200 and a theoretical $\Delta_\theta = 0.2$, the $< 70\%$ power was insufficient to demonstrate a significant predictive ability;

with $\Delta_\theta = 0.1$, large N values were needed. In Scenarios 1, and 2, N = 1600 was necessary to ensure $> 80\%$ power. In Scenario 3, N = 2000 was insufficient to ensure $> 80\%$ power (power was equal to 64%). The power of $\widehat{\Delta}_\theta$ was almost always equal to the power of $\widehat{\gamma}$ and to the one of $\widehat{\beta}_3$ in all settings (Tables 4, 5, and 6).

For the third simulation study, in Scenario 1 and $\Delta_\theta = 0$ (H_0), the α-risk of rejecting H_0 for $\widehat{\Delta}_\theta$ got closer to 5% as N increased; with N = 40 and 1000, the α-risk was equal to 0.0603 and 0.0519, respectively. Similar results were obtained with the interaction coefficient estimation method (Table 7).

For the fourth simulation study, the results show that the α-risk was always close to 5% in case of small differences between the overall risk of event in the two treatment arms, except when the overall risk of event in each treatment arm was low (0.0875 vs. 0.1125, and 0.075 vs. 0.125) where an inflation in the α-risk was observed (e.g. α = 8.69% with N = 5000 and $\rho_{(-1)} = 0.075$ vs. $\rho_{(1)} = 0.125$; Table S3 in Supplemental Material).

4 Application to the PETACC-8 trial

The PETACC-8 trial was led by the *Fédération Francophone de Cancérologie Digestive*. This trial was an open-label, randomized, controlled, multinational phase 3 study that included patients aged 18 to 75 years with pathologically confirmed and resected stage III colon adenocarcinoma (Taieb et al., 2014). The trial compared the efficacy of FOLFOX4 (oxaliplatin, fluorouracil and leucovorin) vs. FOLFOX4 + cetuximab (an inhibitor of EGFR). As patients with *KRAS* exon 2 mutated tumors were found resistant to EGFR antibodies in the metastatic setting (Di Fiore et al., 2007; Lièvre et al., 2008; De Roock et al., 2008), an amendment to the protocol restricted the enrollment to patients with *KRAS* wild-type tumors.

Adding cetuximab to FOLFOX4 did not improve disease-free survival in the intent-to-treat population (hazard ratio: 1.05; 95% CI [0.85; 1.29]) (Taieb et al., 2014). However, the heterogeneity of the response to FOLFOX4 + cetuximab led the investigators to assume that cetuximab could be effective in specific patient subgroups in the per protocol population (N = 1432). The study analyzed the amplification levels of two genes involved in cancer (*DDR2* and *FBXW7*) and considered them as potential treatment selection markers; this restricted the analysis to patients in the per protocol population that had measures of both markers (N = 1068). The capacities of these two genes for treatment selection were assessed in the per protocol population to restrict the analysis to the patients who actually received their treatment. The study outcome was cancer progression or death within 21 months. This delay was a good compromise between the number of censored data and a sufficient time to observe treatment effects (without censoring, the delay would have been too short). The number of censored data was 40; for these, outcomes were imputed (Dmitrienko et al., 2005) (for more details on this imputation, see Supplemental Material). At 21 months, the risk of event occurrence in the FOLFOX4 arm was 0.15 vs. 0.16 in the FOLFOX4 + cetuximab arm.

Concerning the *DDR2* gene (Figure 2), the FOLFOX4 ROC curve was located below the diagonal; in this arm, the risk of event occurrence decreased with the increase of the amplification level of *DDR2*. The FOLFOX4 + cetuximab curve was located above the diagonal; in this arm, the risk of event occurrence increased with the amplification level of *DDR2*. The estimated $\widehat{\Delta}_\theta$ was -0.12 [-0.22; -0.03] and the Wald test p-value = 0.010. The $\widehat{\text{TG}}$ was estimated to 0.04 and the interaction coefficient of the logistic regression model to 18.12 (p-value = 0.040). The estimated $\widehat{\gamma}$ was 0.10 [0.03; 0.16] and the Wald test p-value = 0.003.

Concerning gene *FBXW7* (Figure 3), the FOLFOX4 ROC curve was located above the diagonal, whereas the FOLFOX4 + cetuximab curve was close to the diagonal. Thus, the risk of event occurrence under FOLFOX4 alone increased with the increase of the amplification level of *FBXW7*. The estimated $\widehat{\Delta}_\theta$ was 0.07 [-0.03; 0.16] and the p-value = 0.160. The $\widehat{\text{TG}}$ was estimated to 0.02 and the interaction coefficient of the logistic regression model to -15.22 (p-value = 0.160). The estimated $\widehat{\gamma}$ was 0.04 [-0.03; 0.10] and the Wald test p-value = 0.280.

The $\widehat{\text{TG}}$ and the $\widehat{\gamma}$ concordance measure were higher for the *DDR2* gene than for the *FBXW7* gene, in agreement with the $\widehat{\Delta}_\theta$ results. This means that the capacity of the *DDR2* gene for treatment selection is

higher than that of the *FBXW7* gene. Moreover, the $\widehat{\Delta}_\theta$ of the *DDR2* gene (but not that of the *FBXW7* gene) was significantly different from 0, in agreement with the results from the $\widehat{\gamma}$ concordance index and the interaction coefficient. The *DDR2* gene is thus a treatment selection marker whereas the *FBXW7* gene cannot be considered as a treatment selection marker.

5 Discussion

We present in this article the ABC that may be measured by the Δ_θ metric, a very simple indicator to quantify and test the overall capacity of a quantitative marker for treatment selection when the overall risk of event in each treatment arm is the same. The simulation results showed that the proposed estimation method has good performances. This is reflected by the low mean relative bias in $\widehat{\Delta}_\theta$ ($< 3 \times 10^{-2}$) in all scenarios and all sample sizes, which was comparable to that of the $\widehat{\gamma}$ concordance measure. Furthermore, the mean relative bias of $\widehat{\Delta}_\theta$ was lower than that of the $\widehat{\text{TG}}$ in almost all scenarios and settings, which is explained by the interaction not being linear on the logit scale (except in Scenario 1).

As for $\widehat{\beta}_3$ (the estimator of the interaction coefficient) and the $\widehat{\gamma}$ concordance measure, the power of $\widehat{\Delta}_\theta$ decreased along with the decrease of the sample size and the decrease of the risk of event occurrence. In Scenario 2 and a theoretical $\Delta_\theta = 0.2$, 400 patients were sufficient to reach a power $> 80\%$, whereas in Scenarios 1, and 2 and a theoretical $\Delta_\theta = 0.1$, 1600 patients or more were needed to reach sufficient power. Even higher N values were needed in Scenario 3 because the risk of event occurrence was equal to 0.1 in each arm. Overall, the number of patients needed to achieve sufficient power is in line with the classical need for rather high sample sizes in clinical trials when a test for interaction is performed to detect a treatment selection marker (Janes et al., 2015a). In all settings, the power of $\widehat{\Delta}_\theta$ was comparable to the one of $\widehat{\gamma}$ and to the interaction coefficient estimated by a logistic regression model. These results are interesting because they show that using a non-parametric method - that does not require verifying the complex assumptions required by the parametric methods - would not impact the power of detecting a treatment selection marker, and are in favor of $\widehat{\Delta}_\theta$.

The coverage probability of the asymmetric confidence interval of Δ_θ was always close to 95%, compared to the one of γ that sometimes exceeded 95% (whereas the working model used to calculate the variance of $\widehat{\gamma}$ led to the most optimal variance estimate in the simulations) and to the one of $\widehat{\text{TG}}$ that exceeded 95% in multiple settings. Thus, the use of the asymmetric confidence interval of Δ_θ is recommended. Again, these results are in favor of $\widehat{\Delta}_\theta$.

Three assumptions are necessary to use the proposed Δ_θ metric. The first is that the risk of event in both arms either increases or decreases monotonically over the marker values. This is in fact an assumption similar to but a little more stringent than the assumption of monotonicity of the risk difference between the two treatment arms over the marker values required for other methods proposed to assess a treatment selection marker. The second assumption is that the two ROC curves do not intersect, so that the ABC can be calculated by a simple difference between AUCs. When the ROC curves intersect it is still possible to calculate the ABC using partial AUCs (McClish, 1989), however this approach would need further investigation. The third assumption is that the overall risk of event is the same in the two treatment arms. This assumption can be tested, but small departure from this hypothesis may not be identified by the Chi-squared test. Even in case of small deviations from this assumption, a simulation study showed that the mean $\widehat{\Delta}_\theta$ was always $< 3 \times 10^{-3}$ for markers without treatment selection capacity. The α-risk was still close to 0.05, except for a small overall risk of event (≈ 0.1).

Contrary to the interaction coefficient in a risk model, the Δ_θ metric does not depend on the range of marker values. The γ and the TG indicators have also this property; however, it can be easily demonstrated that their maximum depends on the overall risk of event occurrence in each arm, which is not the case for Δ_θ which is always bounded between -1 and 1. Hence, Δ_θ is an indicator that facilitates comparisons of treatment selection capacities between markers (and between studies). The Δ_θ indicator is thus

useful to detect markers that may be used for treatment decision making, and is complementary with all aforementioned parametric methods.

A binary outcome is necessary to build ROC curves or marker-by-treatment predictiveness curves. However, a frequent problem with binary outcomes is the presence of censored data in long-term trials. In the example of application herein, as there were few censored data at the time of follow-up chosen (40 in a sample of 1068 patients) imputation was not considered to affect significantly the conclusions, but with a longer follow-up the potential impact of imputation may become problematic. To avoid the use of imputation, methods to calculate AUC in the presence of censored data may be used (such as the time-dependent AUC) (Heagerty et al., 2000; Blanche et al., 2013; Li et al., 2018).

The capacities of two markers for treatment selection may be compared by testing the difference in Δ_θs against zero; the standard error of this difference requires the calculation of the covariance between the AUCs of the two markers in each arm. This may use the method developed by DeLong et al. (2011) for correlated ROC curves. But one has to keep in mind that a large $\widehat{\Delta}_\theta$ estimate is not sufficient to detect clinically useful markers and compare them. For example, when the benefit from a treatment increases together with the marker value and the marker-by-treatment predictiveness curves do not intersect, the marker can still be considered as a treatment selection marker because the difference in risks of event is not constant over all marker values. Nevertheless, in this example, one of two treatments is always associated with a lower risk of event; this means that this treatment should be preferred whatever the marker value (and the marker is not useful for treatment choice). Yet, taking into account the mean risks in the target population and the consequences of each treatment strategy (e.g. adverse events), the treatment associated with a higher risk of event occurrence may nevertheless be preferred at marker values where the difference in risks is low if it is associated with fewer adverse events than the other. An extension of the present work would be to include the clinical utility of the treatments in the assessment of treatment selection markers; this requires the estimation of the optimal threshold for treatment allocation that takes into account the clinical utility and the mean risk of event occurrence in each treatment arm (Jund et al., 2005; Subtil and Rabilloud, 2015; Blangero et al., 2019).

In conclusion, the area between ROC curves was able to quantify and test the capacity of a quantitative marker for treatment selection. The method is easy-to-use and complements previous parametric methods when the parametric assumptions cannot be verified.

Acknowledgements The authors thank Dr Philip Robinson (Hospices Civils de Lyon) whose suggestions improved significantly the present manuscript.

Conflict of Interest

The authors have declared no conflict of interest.

Appendix

A.1. Example of a perfect treatment selection marker

In this scenario, two different Δ_θ are computed. One is equal to -1 (Scenario A) and the other to -0.5 (Scenario B). As stated in the manuscript, $\Delta_\theta = -1$ has the same interpretation as $\Delta_\theta = 1$. The marker-by-treatment predictiveness curves that correspond to these scenarios are presented in Figure 4.

In Scenario A ($\Delta_\theta = -1$), the marker-by-treatment predictiveness curves show that below the threshold of 0.5, the risk of event occurrence is equal to 1 in the innovative arm and 0 in the reference one, and that above the threshold of 0.5, the risk of event occurrence is equal to 0 in the innovative arm and 1 in the reference one. Theoretically, no patients is likely to experience the event if the best treatment is given according to the marker values.

In Scenario B ($\Delta_\theta = -0.5$), the ability of the marker to distinguish patients who would and would not experience the event is not maximal. Thus, Δ_θ is not equal to -1 (or 1). The marker remains interesting because the decision to treat changes according to the biomarker value with sometimes high risk-differences between the two treatments. In this case, the patients would not be optimally treated because the treatment allocation process would not prevent the occurrence of the event in all patients. The higher the risk difference between the arms, the farther are the two ROC curves from each other, the higher is Δ_θ, and the stronger is the treatment selection ability of the marker.

A.2. Validity of the Δ_θ indicator

First, let us demonstrate that when a marker has no capacity for treatment selection then the Δ_θ indicator is equal to 0. When a marker has no capacity for treatment selection:

$$\delta(v) = \mathrm{P}(E=1|T=-1,V=v) - \mathrm{P}(E=1|T=1,V=v) = \rho_{(-1)} - \rho_{(1)} = 0 \qquad \forall v$$

With the assumption (1) that the overall risk in each treatment arm is equal, $\rho_{(-1)} - \rho_{(1)} = 0$.

The expression of the Δ_θ indicator is

$$\Delta_\theta = \frac{\int_{-\infty}^{+\infty} F(v) \times \delta(v)\, dF(v)}{\rho(1-\rho)}$$

It is easy to see that when a marker has no capacity for treatment selection then the numerator is equal to 0. So when a marker has no capacity for treatment selection $\Delta_\theta = 0$.

Now, let us demonstrate that when $\Delta_\theta = 0$, then a marker has no capacity for treatment selection. With the assumption that the ROC curves do not intersect, $\Delta_\theta = 0$ means that the ROC curves for the two treatment arms overlap. This means that for the infinity of couples $(c_{(-1)}, c_{(1)})$ that correspond to the set Ω of marker values for which the ROC curves overlap, the following systems must hold:

$$\begin{cases} \mathrm{P}(V > c_{(-1)}|E=1,T=-1) = \mathrm{P}(V > c_{(1)}|E=1,T=1) \\ \mathrm{P}(V > c_{(-1)}|E=0,T=-1) = \mathrm{P}(V > c_{(1)}|E=0,T=1) \end{cases} \qquad \forall (c_{(-1)}, c_{(1)}) \in \Omega$$

$$\begin{cases} \mathrm{P}(V \leq c_{(-1)}|E=1,T=-1) = \mathrm{P}(V \leq c_{(1)}|E=1,T=1) \\ \mathrm{P}(V \leq c_{(-1)}|E=0,T=-1) = \mathrm{P}(V \leq c_{(1)}|E=0,T=1) \end{cases} \qquad \forall (c_{(-1)}, c_{(1)}) \in \Omega \qquad (4)$$

For a given couple $(c_{(-1)}, c_{(1)})$, the sensitivity and specificity in one treatment arm must be equal to those of the other treatment arm.

It is also important to respect the randomization constraint that is assumed in a clinical trial context (equation (3)). This can be expressed as:

$$\mathrm{P}(V \leq c|T=-1) = \mathrm{P}(V \leq c|T=1) \qquad \forall c$$

This expression can be rewritten according to assumption (1) :

$$\begin{aligned} &\rho[\mathrm{P}(V \leq c|E=1,T=-1) - \mathrm{P}(V \leq c|E=1,T=1)] \\ &\quad - (1-\rho)[\mathrm{P}(V \leq c|E=0,T=1) - \mathrm{P}(V \leq c|E=0,T=-1)] = 0 \end{aligned} \qquad \forall c$$

This equation must be true at the point $c_{(-1)}$:

$$\begin{aligned} &\rho[\mathrm{P}(V \leq c_{(-1)}|E=1,T=-1) - \mathrm{P}(V \leq c_{(-1)}|E=1,T=1)] \\ &\quad - (1-\rho)[\mathrm{P}(V \leq c_{(-1)}|E=0,T=1) - \mathrm{P}(V \leq c_{(-1)}|E=0,T=-1)] = 0 \end{aligned}$$

Replacing by the information provided in system (4), it is possible to write:

$$\begin{aligned}\rho[&\mathrm{P}(V \leq c_{(1)}|E=1,T=1) - \mathrm{P}(V \leq c_{(-1)}|E=1,T=1)] \\ &- (1-\rho)[\mathrm{P}(V \leq c_{(-1)}|E=0,T=1) - \mathrm{P}(V \leq c_{(1)}|E=0,T=1)] = 0\end{aligned}$$

If $c_{(-1)} > c_{(1)}$, the sensitivity in the innovative treatment for $c_{(-1)}$ is lower than for $c_{(1)}$, and the specificity of the innovative treatment for $c_{(-1)}$ is greater than for $c_{(1)}$; thus, the left part of the equation is strictly negative and the equation cannot be true.

If $c_{(-1)} < c_{(1)}$, the sensitivity in the innovative treatment for $c_{(-1)}$ is greater than for $c_{(1)}$, and the specificity of the innovative treatment for $c_{(-1)}$ is lower than for $c_{(1)}$; thus, the left part of the equation is strictly positive and the equation again cannot be true.

So the equation is only true when $c_{(-1)} = c_{(1)} = c$. It is then possible ro write again the system (4) including this constraint:

$$\begin{cases}\mathrm{P}(V \leq c)|E=1,T=-1) = \mathrm{P}(V \leq c|E=1,T=1) \\ \mathrm{P}(V \leq c)|E=0,T=-1) = \mathrm{P}(V \leq c|E=0,T=1)\end{cases} \qquad \forall c$$

Using the Bayes theorem, this implies that

$$\mathrm{P}(E=1|T=-1,V \leq c) = \mathrm{P}(E=1|T=1,V \leq c) \qquad \forall c$$

Where the result comes from assumption (1) and the randomization constraint. It follows that

$$\mathrm{P}(E=1|T=-1,V=c) = \mathrm{P}(E=1|T=1,V=c) = \rho_{(-1)} - \rho_{(1)} = 0 \qquad \forall c$$

Which is the definition of a marker without capacity for treatment selection given in equation (2).

References

Ballman, K. V. (2015). Biomarker: Predictive or Prognostic? *Journal of Clinical Oncology* **33**, 3968–3971.

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal* **55**, 687–704.

Blangero, Y., Rabilloud, M., Ecochard, R., and Subtil, F. (2019). A Bayesian method to estimate the optimal threshold of a marker used to select patients' treatment. *Statistical Methods in Medical Research*. DOI: 10.1177/0962280218821394

Byar, D. P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine* **4**, 255–263.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845.

De Roock, W., Piessevaux, H., De Schutter, J., Janssens, M., De Hertogh, G., Personeni, N., Biesmans, B., Van Laethem, J. L., Peeters, M., Humblet, Y., Van Cutsem, E., and Tejpar, S. (2008). KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. *Annals of Oncolofy* **19**, 508–515.

Di Fiore, F., Blanchard, F., Charbonnier, F., Le Pessot, F., Lamy, A., Galais, M. P., Bastit, L., Killian, A., Sesboüé, R., Tuech, J. J., Queuniet, A. M., Paillot, B., Sabourin, J. C., Michot, F., Michel, P., and Frebourg, T. (2007). Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. *British Journal of Cancer* **96**, 1166–1169.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. SAS Institute, Cary.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton.

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19**, 293–325.

Huang, Y., Gilbert, P. B., Janes, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* **68**, 687–696.

Italiano, A. (2011). Prognostic or Predictive? It's time to get back to definitions! *Journal of Clinical Oncology* **29**, 4718.

Janes, H., Brown, M. D., Huang, Y., Pepe, M. S. (2014a). An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics* **10**, 99–121.

Janes, H., Brown, M. D., Pepe, M. S. (2015a). Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in Medicine* **34**, 3503–3515.

Janes, H., Pepe, M. S., Bossuyt, P. M., and Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine* **154**, 253–259.

Janes, H., Pepe, M. S., and Huang, Y. (2014b). A framework for evaluating markers used to select patient treatment. *Medical Decision Making* **34**, 159–167.

Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., and Heagerty, P. J. (2015b). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute* **107**. DOI: 10.1093/jnci/djv157

Jund, J., Rabilloud, M., Wallon, M., and Ecochard, R. (2005). Methods to estimate the optimal threshold for normally or log-normally distributed biological tests. *Medical Decision Making* **25**, 406–415.

Li, L., Greene, T., and Hu, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research* **27**, 2264–2278.

Lièvre, A., Bachet, J. B., Boige, V., Cayre, A., Le Corre, D., Buc, E., Ychou, M., Bouché, O., Landi, B., Louvet, C., André, T., Bibeau, F., Diebold, M. D., Rougier, P., Ducreux, M., Tomasic, G., Emile, J. F., Penault-Llorca, F., and Laurent-Puig, P. (2008). KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *Journal of Clinical Oncology* **26**, 374–379.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

Metz, C. E., and Pan, X. (1999). "Proper" binormal ROC curves: Theory and maximum-likelihood estimation. *Journal of Mathematical Psychology* **43**, 1–33.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Skougaard, K., Nielsen, D., Jensen, B. V., Pfeiffer, P., and Hendel, H. W. (2016). Early (18)F-FDG-PET/CT as a predictive marker for treatment response and survival in patients with metastatic colorectal cancer treated with irinotecan and cetuximab. *Acta Oncology* **55**, 1175–1182.

Song, X., and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60**, 874–883.

Subtil, F., and Rabilloud, M. (2015). An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. *Journal of Clinical Epidemiology* **68**, 752–759.

Taieb, J., Tabernero, J., Mini, E., Subtil, F., Folprecht, G., Van Laethem, J. L., Thaler, J., Bridgewater, J., Petersen, L. N., Blons, H., Collette, L., Van Cutsem, E., Rougier, P., Salazar, R., Bedenne, L., Emile, J. F., Laurent-Puig, P., and Lepage, C. (2014). Oxaliplatin, fluorouracil, and leucovorin with or without cetuximab in patients with resected stage III colon cancer (PETACC-8): an open-label, randomised phase 3 trial. *The Lancet Oncology* **15**, 862–873.

Viallon, V., and Latouche, A. (2011). Discrimination measures for survival outcomes: Connection between the AUC and the predictiveness curve. *Biometrical Journal* **53**, 217–236.

Vickers, A. J., Kattan, M. W., and Sargent, D. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* **8**, 14.

Weidhaas, J. B., Harris, J., Schaue, D., Chen, A. M., Chin, R., Axelrod, R., El-Naggar, A. K., Singh, A. K., Galloway, T. J., Raben, D., Wang, D., Matthiesen, C., Avizonis, V. N., Manon, R. R., Yumen, O., Nguyen-Tan, P. F., Trotti, A., Skinner, H., Zhang, Q., Ferris, R. L., Sidransky, D., and Chung, C. H. (2016). The KRAS-Variant and Cetuximab Response in Head and Neck Squamous Cell Cancer: A Secondary Analysis of a Randomized Clinical Trial. *JAMA Oncology* **3**, 483–491.

Zhang, Z., Nie, L., Soon, G., and Liu, A. (2014). The Use of Covariates and Random Effects in Evaluating Predictive Biomarkers Under a Potential Outcome Framework. *Annals of Applied Statistics* **8**, 2336–2355.

Zhang, Z., Ma, S., Nie, L., and Soon, G. (2017). A quantitative concordance measure for comparing and combining treatment selection markers. *The International Journal of Biostatistics* **13**. DOI: 10.1515/ijb-2016-0064

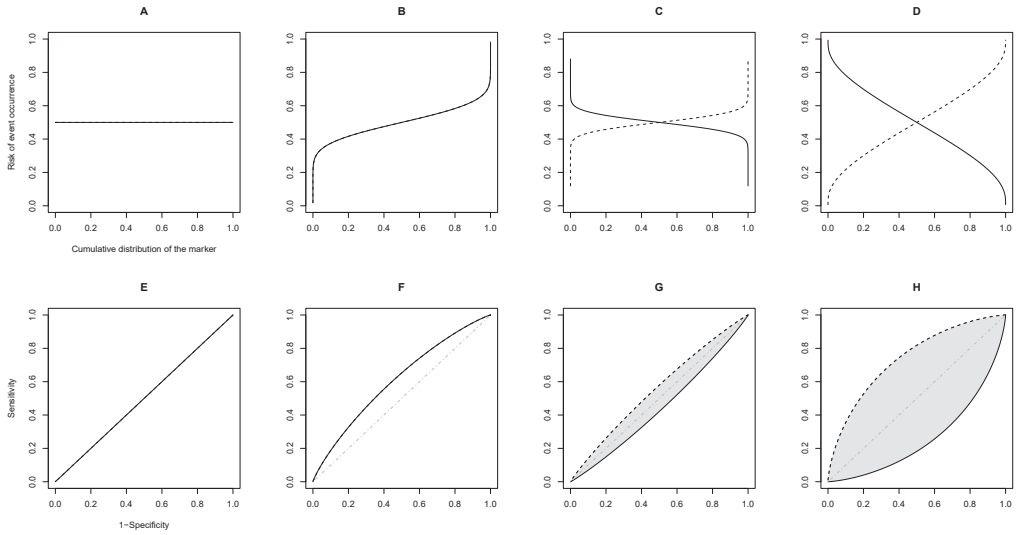


Figure 1 Marker-by-treatment predictiveness curves of four markers, and their corresponding ROC curves (Dotted line: innovative treatment; solid line: reference treatment; shaded area: area between ROC curves).

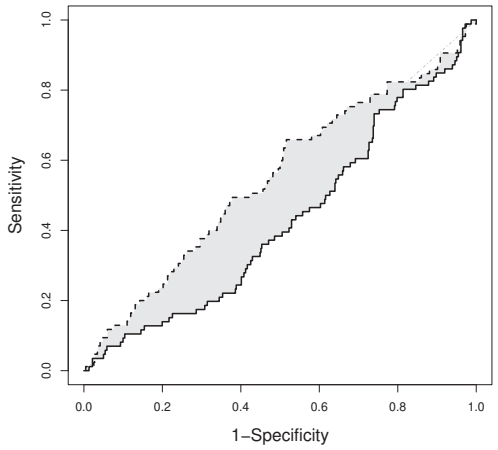


Figure 2 ROC curves associated with the *DDR2* gene (Dotted line: FOLFOX4+Cetuximab treatment; Solid line:FOLFOX4 treatment).

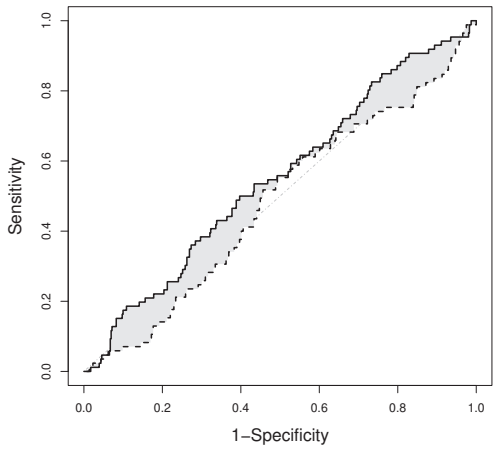


Figure 3 ROC curves associated with the *FBXW7* gene (Dotted line: FOLFOX4+Cetuximab treatment; Solid line:FOLFOX4 treatment).

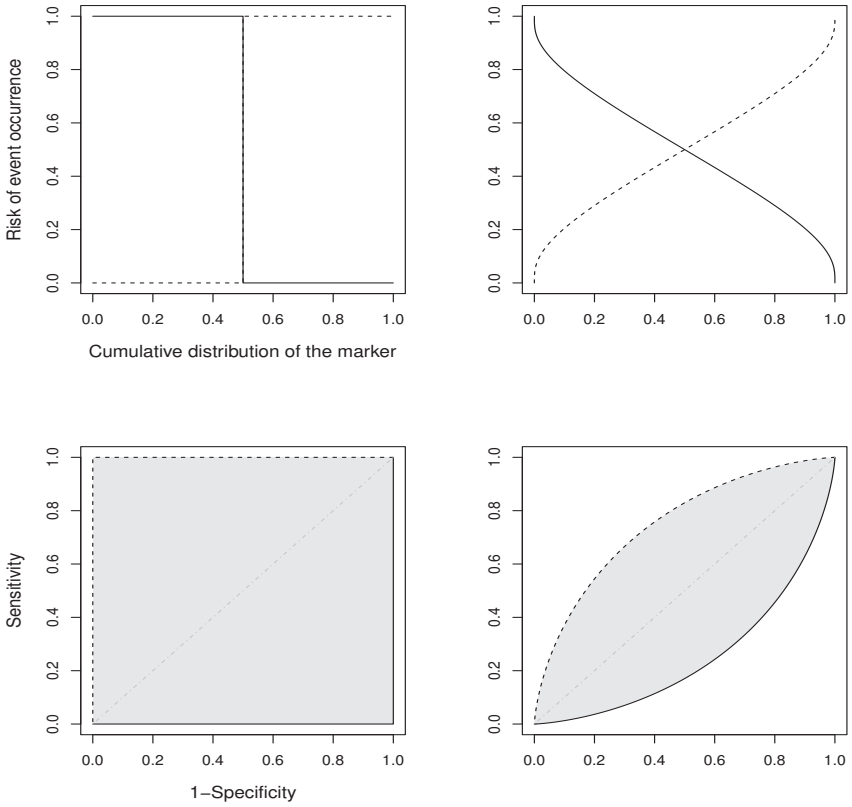


Figure 4 Marker-by-treatment predictiveness curves of two markers, and their corresponding ROC curves (Dotted line: innovative treatment; solid line: reference treatment; shaded area: area between ROC curves).

Table 1 Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of Δ_θ (Scenario 1)

		Symmetric		Asymmetric	
Δ_θ	N	CP	WCI	CP	WCI
0.95	200	0.9081	0.07	0.9474	0.07
	400	0.9339	0.05	0.9519	0.05
	1000	0.9416	0.03	0.9498	0.03
	1600	0.9484	0.02	0.9517	0.02
	2000	0.9441	0.02	0.9493	0.02
0.6	200	0.9439	0.24	0.9507	0.24
	400	0.9482	0.17	0.9519	0.17
	1000	0.9471	0.11	0.9480	0.11
	1600	0.9491	0.09	0.9500	0.09
	2000	0.9515	0.08	0.9512	0.08
0.4	200	0.9466	0.29	0.9526	0.29
	400	0.9495	0.20	0.9526	0.20
	1000	0.9462	0.13	0.9484	0.13
	1600	0.9493	0.10	0.9495	0.10
	2000	0.9503	0.09	0.9508	0.09
0.2	200	0.9489	0.31	0.9537	0.31
	400	0.9489	0.22	0.9509	0.22
	1000	0.9461	0.14	0.9480	0.14
	1600	0.9482	0.11	0.9483	0.11
	2000	0.9500	0.10	0.9502	0.10
0.1	200	0.9493	0.32	0.9532	0.32
	400	0.9481	0.23	0.9499	0.23
	1000	0.9481	0.14	0.9480	0.14
	1600	0.9476	0.11	0.9483	0.11
	2000	0.9498	0.10	0.9497	0.10

Δ_θ: area between ROC curves; N: sample size; CP: coverage probability; WCI: mean width of the confidence interval

Table 2 Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of Δ_θ (Scenario 2)

		Symmetric		Asymmetric	
Δ_θ	N	CP	WCI	CP	WCI
0.6	200	0.9414	0.24	0.9511	0.24
	400	0.9484	0.17	0.9526	0.17
	1000	0.9486	0.11	0.9482	0.11
	1600	0.9468	0.09	0.9481	0.09
	2000	0.9462	0.08	0.9469	0.08
0.4	200	0.9487	0.32	0.9541	0.32
	400	0.9473	0.22	0.9512	0.22
	1000	0.9505	0.14	0.9509	0.14
	1600	0.9505	0.11	0.9504	0.11
	2000	0.9490	0.10	0.9594	0.10
0.2	200	0.9481	0.36	0.9533	0.36
	400	0.9472	0.25	0.9496	0.25
	1000	0.9507	0.16	0.9515	0.16
	1600	0.9512	0.13	0.9519	0.13
	2000	0.9514	0.11	0.9528	0.11
0.1	200	0.9500	0.37	0.9554	0.37
	400	0.9474	0.26	0.9503	0.26
	1000	0.9542	0.16	0.9561	0.16
	1600	0.9501	0.13	0.9501	0.13
	2000	0.9490	0.12	0.9503	0.12

Δ_θ: area between ROC curves; N: sample size; CP: coverage probability; WCI: mean width of the confidence interval

Table 3 Coverage probability and mean width of the 95% symmetric and asymmetric confidence intervals of Δ_θ (Scenario 3)

		Symmetric		Asymmetric	
Δ_θ	N	CP	WCI	CP	WCI
0.4	200	0.9325	0.43	0.9458	0.43
	400	0.9417	0.30	0.9477	0.30
	1000	0.9482	0.19	0.9495	0.19
	1600	0.9512	0.15	0.9518	0.15
	2000	0.9505	0.14	0.9529	0.14
0.2	200	0.9334	0.51	0.9449	0.50
	400	0.9419	0.36	0.9459	0.36
	1000	0.9468	0.23	0.9493	0.23
	1600	0.9529	0.18	0.9536	0.18
	2000	0.9485	0.16	0.9501	0.16
0.1	200	0.9335	0.53	0.9446	0.52
	400	0.9424	0.37	0.9478	0.37
	1000	0.9483	0.24	0.9504	0.24
	1600	0.9535	0.19	0.9547	0.19
	2000	0.9496	0.17	0.9508	0.17

Δ_θ: area between ROC curves; N: sample size; CP: coverage probability; WCI: mean width of the confidence interval

Table 4 Performance of the treatment selection indicators (Scenario 1)

		$\widehat{\Delta}_\theta$			$\widehat{\gamma}$			$\widehat{\text{TG}}$		$\widehat{\beta}_3$
Δ_θ	N	MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.6	200	0.00102	0.9507	1	0.00041	0.9367	1	0.01242	0.9366	1
	400	-0.00040	0.9519	1	-0.00079	0.9436	1	0.00498	0.9442	1
	1000	-0.00104	0.9480	1	-0.00120	0.9454	1	0.00105	0.9452	1
	1600	-0.00063	0.9500	1	-0.00071	0.9486	1	0.00059	0.9473	1
	2000	-0.00015	0.9512	1	-0.00023	0.9507	1	0.00084	0.9477	1
0.4	200	0.00246	0.9526	0.9990	0.00087	0.9424	0.9990	0.01626	0.9394	0.9991
	400	-0.00073	0.9526	1	-0.00153	0.9461	1	0.00605	0.9448	1
	1000	-0.00190	0.9484	1	-0.00226	0.9460	1	0.00077	0.9468	1
	1600	-0.00110	0.9495	1	-0.00130	0.9494	1	0.00050	0.9479	1
	2000	-0.00031	0.9508	1	-0.00048	0.9502	1	0.00103	0.9474	1
0.2	200	0.00669	0.9537	0.7062	0.00469	0.9435	0.7195	0.02378	0.9473	0.7170
	400	-0.00116	0.9509	0.9377	-0.00225	0.9462	0.9378	0.00646	0.9462	0.9450
	1000	-0.00417	0.9480	0.9996	-0.00463	0.9457	0.9996	-0.00104	0.9493	0.9999
	1600	-0.00227	0.9483	1	-0.00259	0.9483	1	-0.00049	0.9479	1
	2000	-0.00057	0.9502	1	-0.00087	0.9497	1	0.00101	0.9488	1
0.1	200	0.01491	0.9532	0.2450	0.01260	0.9432	0.2585	0.11453	0.9696	0.2473
	400	-0.00220	0.9499	0.4182	-0.00331	0.9462	0.4265	0.02345	0.9718	0.4267
	1000	-0.00838	0.9480	0.7840	-0.00892	0.9469	0.7858	-0.00475	0.9502	0.8042
	1600	-0.00463	0.9483	0.9390	-0.00498	0.9469	0.9400	-0.00283	0.9482	0.9456
	2000	-0.00118	0.9497	0.9739	-0.00153	0.9494	0.9744	0.00063	0.9495	0.9795

Δ_θ: area between ROC curves; N: sample size; MRB: mean relative bias; CP: coverage probability; β_3: interaction coefficient

Table 5 Performance of the treatment selection indicators (Scenario 2)

		$\widehat{\Delta}_\theta$			$\widehat{\gamma}$			$\widehat{\text{TG}}$		$\widehat{\beta}_3$
Δ_θ	N	MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.6	200	0.00023	0.9511	1	-0.00158	0.9765	1	-0.03993	0.9286	0.9998
	400	-0.00088	0.9526	1	-0.00244	0.9819	1	-0.04663	0.9038	1
	1000	-0.00075	0.9482	1	-0.00179	0.9788	1	-0.05005	0.7978	1
	1600	0.00113	0.9481	1	0.00161	0.9821	1	-0.04830	0.7091	1
	2000	0.00084	0.9469	1	0.07447	0.9824	1	-0.04890	0.6445	1
0.4	200	0.00168	0.9541	0.9971	-0.00078	0.9619	0.9970	-0.00747	0.9434	0.9980
	400	0.00206	0.9512	1	0.00046	0.9651	1	-0.01416	0.9479	1
	1000	-0.00067	0.9509	1	-0.00092	0.9670	1	-0.02141	0.9467	1
	1600	-0.00013	0.9504	1	-0.00005	0.9664	1	-0.02103	0.9388	1
	2000	0.00109	0.9504	1	0.00075	0.9662	1	-0.02081	0.9363	1
0.2	200	0.02271	0.9533	0.6048	0.01992	0.9490	0.6102	0.03718	0.9561	0.6184
	400	0.00763	0.9496	0.8700	0.00634	0.9488	0.8727	0.00863	0.9439	0.8836
	1000	-0.00082	0.9515	0.9985	-0.00166	0.9545	0.9986	-0.00515	0.9511	0.9986
	1600	-0.00207	0.9519	1	-0.00289	0.9555	1	-0.00747	0.9485	1
	2000	0.00300	0.9528	1	0.00267	0.9556	1	-0.00129	0.9485	1
0.1	200	0.03636	0.9554	0.2005	0.03423	0.9489	0.2084	0.17649	0.9671	0.2013
	400	0.02256	0.9503	0.3398	0.02148	0.9474	0.3467	0.05647	0.9700	0.3444
	1000	-0.00036	0.9561	0.6605	-0.00088	0.9549	0.6627	0.00176	0.9573	0.6798
	1600	-0.00885	0.9501	0.8513	-0.00905	0.9514	0.8504	-0.00991	0.9496	0.8634
	2000	-0.00157	0.9503	0.9180	-0.00252	0.9500	0.9187	0.00144	0.9474	0.9283

Δ_θ: area between ROC curves; N: sample size; MRB: mean relative bias; CP: coverage probability; β_3: interaction coefficient

Table 6 Performance of the treatment selection indicators (Scenario 3)

		$\widehat{\Delta}_\theta$			$\widehat{\gamma}$			$\widehat{\text{TG}}$		$\widehat{\beta}_3$
Δ_θ	N	MRB	CP	Power	MRB	CP	Power	MRB	CP	Power
0.4	200	0.00580	0.9458	0.9406	-0.00215	0.9848	0.8764	0.00434	0.9277	0.9516
	400	0.00296	0.9477	0.9989	0.00061	0.9854	0.9951	-0.00481	0.9390	0.9997
	1000	-0.00243	0.9495	1	-0.00542	0.9859	1	-0.01636	0.9482	1
	1600	0.00058	0.9518	1	0.00101	0.9879	1	-0.01227	0.9474	1
	2000	0.00005	0.9529	1	-0.00093	0.9868	1	-0.01496	0.9449	1
0.2	200	0.01893	0.9449	0.3624	0.01393	0.9580	0.3244	0.05878	0.9491	0.3471
	400	0.00556	0.9459	0.5880	0.00331	0.9606	0.5584	0.01065	0.9501	0.6045
	1000	0.00157	0.9493	0.9339	0.00054	0.9626	0.9248	-0.00024	0.9476	0.9476
	1600	-0.00050	0.9536	0.9939	-0.00068	0.9657	0.9926	-0.00271	0.9488	0.9961
	2000	0.00190	0.9501	0.9981	0.00171	0.9627	0.9980	-0.00072	0.9484	0.9992
0.1	200	0.02403	0.9446	0.1327	0.02016	0.9473	0.1172	0.37949	0.9547	0.1158
	400	0.01717	0.9478	0.1942	0.01630	0.9493	0.1841	0.15611	0.9623	0.1921
	1000	0.00072	0.9504	0.3800	0.00013	0.9530	0.3713	0.02104	0.9723	0.3910
	1600	0.00049	0.9547	0.5482	0.00097	0.9573	0.5416	0.00438	0.9675	0.5648
	2000	0.00228	0.9508	0.6514	0.00199	0.9536	0.6452	0.00613	0.9546	0.6704

Δ_θ: area between ROC curves; N: sample size; MRB: mean relative bias; CP: coverage probability; β_3: interaction coefficient

Table 7 α-risk under H_0 (third simulation study)

		$\widehat{\Delta}_\theta$	$\widehat{\beta}_3$
Δ_θ	N	α	α
0	40	0.0603	0.0465
	100	0.0503	0.0459
	200	0.0500	0.0506
	400	0.0527	0.0512
	1000	0.0519	0.0487

Δ_θ: area between ROC curves; N: sample size

2.5.2 Principaux résultats de l'article

Les résultats des simulations montrent que la méthode d'estimation de Δ_θ a de bonnes performances. Cela est notamment reflété par le biais moyen relatif au plus inférieur à 3×10^{-2} dans tous les scénarios et tailles d'échantillon, et qui est toujours comparable à celui de l'estimateur de l'indice de concordance γ. De plus ce biais relatif moyen est inférieur à celui de l'estimateur du TG dans presque tous les scénarios envisagés ; cela est expliqué entre autres par l'interaction non linéaire sur l'échelle du logit dans les Scénarios 2 et 3.

En termes de puissance, celle de $\widehat{\Delta}_\theta$ est comparable à celle de $\widehat{\gamma}$, mais surtout à celle du coefficient d'interaction estimé par un modèle paramétrique. Ainsi, l'approche non paramétrique proposée n'entraîne pas une perte de puissance par rapport aux approches paramétriques classiques, sans nécessiter la vérification complexe des hypothèses de ces modèles. La probabilité de couverture de l'intervalle de confiance asymétrique de Δ_θ est toujours proche de 95 %, contrairement à celle de γ et du TG dans certains cas. Par ailleurs, Δ_θ est une métrique bornée entre 0 et 1, et indépendante des risques moyens d'événement dans les bras de traitement, ce qui facilite la comparaison des résultats entre les études. Tous ces résultats sont en faveur de l'utilisation de Δ_θ.

Concernant l'analyse des niveaux d'amplification des gènes *DDR2* et *FBXW7* dans le cadre de l'essai PETACC-8, tous les estimateurs concordent et permettent de conclure que le gène *DDR2* possède une capacité prédictive statistiquement significative.

2.6 Compléments à l'article

2.6.1 Calcul de Δ_θ à partir de distributions théoriques de marqueur

Les simulations de l'article ont nécessité le calcul des valeurs théoriques de Δ_θ selon les scénarios envisagés. Ces calculs sont détaillés ci-après.

Le cas des courbes ROC binormales

Soient $X^{(zY)}$ le marqueur dans le bras $Z = z$ pour les patients qui développent l'évènement d'intérêt ($Y = 1$), et $X^{(z\bar{Y})}$ le marqueur dans le bras $Z = z$ pour les patients qui ne développent pas l'évènement d'intérêt ($Y = 0$). Supposons que les valeurs de marqueur dans les quatre groupes possibles ($1Y$, $1\bar{Y}$, $0Y$, et $0\bar{Y}$) suivent des lois normales :

$$
\begin{aligned}
X^{(1Y)} &\sim \mathcal{N}\left(\mu^{(1Y)}, (\sigma^{(1Y)})^2\right), \\
X^{(1\bar{Y})} &\sim \mathcal{N}\left(\mu^{(1\bar{Y})}, (\sigma^{(1\bar{Y})})^2\right), \\
X^{(0Y)} &\sim \mathcal{N}\left(\mu^{(0Y)}, (\sigma^{(0Y)})^2\right), \\
X^{(0\bar{Y})} &\sim \mathcal{N}\left(\mu^{(0\bar{Y})}, (\sigma^{(0\bar{Y})})^2\right).
\end{aligned}
$$

L'expression des AUCs issues de courbes ROC binormales permet de calculer de manière analytique Δ_θ comme :

$$\begin{aligned}\Delta_\theta &= \theta_1 - \theta_0 \\ &= \Phi\left(\frac{a_1}{\sqrt{1+b_1^2}}\right) - \Phi\left(\frac{a_0}{\sqrt{1+b_0^2}}\right),\end{aligned}$$

avec $\Phi(.)$ la fonction de répartition d'une loi normale centrée réduite, $a_z = \frac{\mu^{(zY)} - \mu^{(z\bar{Y})}}{\sigma^{(zY)}}$, et $b_z = \frac{\sigma^{(z\bar{Y})}}{\sigma^{(zY)}}$ pour $Z = z$ (Pepe, 2003).

Les autres cas

Lorsque les valeurs de marqueur dans les différents groupes de patient ne suivent pas des lois normales, il n'existe pas forcément de formule analytique de chaque AUC. En revanche il est possible de calculer de manière numérique Δ_θ pour des distributions fixées de marqueur dans chaque groupe :

$$\Delta_\theta = \int_{-\infty}^{+\infty}\int_{c}^{+\infty} f^{(1Y)}(x) f^{(1\bar{Y})}(c)\, dxdc - \int_{-\infty}^{+\infty}\int_{c}^{+\infty} f^{(0Y)}(x) f^{(0\bar{Y})}(c)\, dxdc,$$

où $f^{(zY)}(.)$ est la densité de probabilité du marqueur pour $Z = z$ et $Y = 1$, et $f^{(z\bar{Y})}(.)$ la densité de probabilité du marqueur pour $Z = z$ et $Y = 0$.

2.6.2 Extension de la méthode à la prise en compte des censures

Comme la survenue de l'évènement n'est évaluée qu'après un délai d'observation fixé, il est possible que l'on n'ait pas l'information pour tous les patients. En présence de données censurées, il est aujourd'hui bien connu que l'estimateur naïf de l'AUC est biaisé, et ce même lorsque la distribution des censures est uniforme sur l'intervalle de temps $[0,t]$. Soient T le temps jusqu'à l'évènement d'intérêt étudié, et C le temps jusqu'à la censure, alors l'expression du biais a déjà été démontrée dans la littérature (Blanche et al., 2013b), et s'exprime de la manière suivante :

$$\widehat{\theta}_{Naif}(t) \xrightarrow{a.s.} \theta(t) \times \frac{\mathrm{P}(T_i \leq C_i, C_j > t | T_i \leq t, T_j > t, X_i > X_j)}{\mathrm{P}(T_i \leq C_i, C_j > t | T_i \leq t, T_j > t)}.$$

A partir du moment où T dépend de X, il existe un biais. Cependant, les auteurs ont démontré qu'en pratique ce biais demeurait assez faible. Plusieurs méthodes de correction du biais ont été proposées, deux d'entre elles ont été évaluées dans cette thèse pour estimer Δ_θ : la méthode de correction par Inverse Probability of Censoring Weighting (IPCW) (Hung and Chiang, 2010), et la méthode de Li et al. (2018).

Inverse Probability of Censoring Weighting

Les méthodes de type Inverse Probability Weighting sont courantes, et largement développées dans la littérature pour des cas autres que la correction du biais de l'estimateur de l'AUC en présence de données censurées (Robins et al., 1994; Hernan and Robins, 2006). Leur objectif est de pondérer les patients par l'inverse d'une probabilité qui sera définie selon le contexte où elles sont appliquées. Par exemple, dans le cas de l'IPCW, les patients restant dans l'analyse sont pondérés par l'inverse de leur probabilité de ne pas être censurés. Son application spécifique à l'AUC a été proposée et étudiée par plusieurs auteurs (Hung and Chiang, 2010; Blanche et al., 2013a). Soient $Y = \min(T,C)$ le temps observé dans les données jusqu'à l'évènement ou la censure, $\delta = \mathbb{1}(Y = T)$ l'indicatrice de censure, $S_T(t) = \mathrm{P}(T > t)$ la fonction de survie, $S_C(t) = \mathrm{P}(C > t)$ la fonction de « survie » des censures, et $S_Y(t) = \mathrm{P}(Y > t)$ la fonction de survie observée.

La formule de l'AUC est modifiée en pondérant les patients présentant l'évènement dans l'intervalle $[0,t]$, ainsi que les patients n'ayant pas encore développé l'évènement :

$$\theta_{IPCW}(t) = \frac{\mathbf{E}\left[\delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\mathbb{1}(X_i > X_j)\frac{1}{S_C(Y_i|X_i)S_C(t|X_i)}\right]}{\mathbf{E}\left[\delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\frac{1}{S_C(Y_i|X_i)S_C(t|X_i)}\right]}.$$

L'estimateur de l'AUC s'écrit alors de la manière suivante :

$$\widehat{\theta}_{IPCW}(t) = \frac{\sum_{i\neq j} \delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\mathbb{1}(X_i > X_j)\frac{1}{\widehat{S}_C(Y_i|X_i)\widehat{S}_C(t|X_i)}}{\sum_{i\neq j} \delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\frac{1}{\widehat{S}_C(Y_i|X_i)\widehat{S}_C(t|X_i)}}.$$

Sous l'hypothèse que C est indépendant de (T,X), $S_C(t|x)$ se réduit à $S_C(t)$, et l'estimateur se reformule de la manière suivante :

$$\widehat{\theta}_{IPCW}(t) = \frac{\sum_{i\neq j} \delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\mathbb{1}(X_i > X_j)\widehat{S}_C(Y_i)^{-1}}{n(n-1)\widehat{S}_Y(t)(1-\widehat{S}_T(t))},$$

avec $\widehat{S}_Y(t)$ l'estimateur empirique de $S_Y(t)$, et $\widehat{S}_T(t)$ et $\widehat{S}_C(t)$ les estimateurs de Kaplan-Meier de $S_T(t)$ et $S_C(t)$ respectivement.

Alors que certains auteurs proposent d'utiliser des méthodes bootstrap pour obtenir un intervalle de confiance de cet estimateur (Blanche et al., 2013b), d'autres ont proposé une expression asymptotique de la variance de l'estimateur permettant alors d'estimer directement un intervalle de confiance (Hung and Chiang, 2010).

Méthode par bootstrap

L'objectif du bootstrap est d'approcher par simulation la distribution d'un estimateur lorsque sa distribution est inconnue, ou bien qu'elle ne peut tout simplement pas être approximée par une loi normale (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). A partir des données initiales, le principe est de rééchantillonner avec remise dans le but de construire K échantillons bootstrap. L'estimateur est alors appliqué à chacun de ces K échantillons bootstrap, permettant

d'obtenir K estimations. Ces K estimations vont former la distribution boostrap à partir de laquelle plusieurs méthodes peuvent être appliquées.

La méthode boostrap la plus commune est le bootstrap percentile. A partir de la distrbution bootstrap, il est possible de construire un intervalle de confiance à $(1-\alpha)$ % de l'estimateur en utilisant les quantiles $\alpha/2$ et $1-\alpha/2$. C'est principalement cette méthode qui sera appliquée dans le cadre de cette thèse.

Variance asymptotique de l'estimateur

Dans leur article, Hung and Chiang (2010) propose une expression asymptotique de la variance de l'estimateur de l'AUC en présence de censures. Si on pose les notations suivantes :

$$h_{tij} = \delta_i \mathbb{1}(Y_i \leq t, Y_j > t)\mathbb{1}(X_i > X_j) S_C(Y_i)^{-1},$$

$$h_t = \mathbf{E}(h_{tij}),$$

$$M_{T_i}(t) = \delta_i \mathbb{1}(Y_i \leq t) + \int_0^t \mathbb{1}(Y_i \geq u)\, du\{\log[S_T(u)]\},$$

$$M_{C_i}(t) = \delta_i \mathbb{1}(Y_i \leq t) + \int_0^t \mathbb{1}(Y_i \geq u)\, du\{\log[S_C(u)]\},$$

Hung and Chiang (2010) démontrent que $\sqrt{n}[\widehat{\theta}_{IPCW}(t) - \theta_{IPCW}(t)]$ converge vers une loi normale centrée sur 0 et de variance σ_t^2. Un estimateur de σ_t^2 est donné par :

$$\widehat{\sigma}_t^2 = n^{-1} \sum_{i=1}^{n} \widehat{\mathrm{IF}}_i(t)^2,$$

où $\widehat{\mathrm{IF}}_i(.)$ est la fonction d'influence empirique évaluée pour l'individu i et exprimée comme :

$$\widehat{\mathrm{IF}}_i(t) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \left[\widehat{\Psi}_{ijk}(t) + \widehat{\Psi}_{jik}(t) + \widehat{\Psi}_{jki}(t) \right],$$

et $\widehat{\Psi}_{ijk}(t)$ est calculé en utilisant les estimateurs de $\widehat{h}_t$, $\widehat{h}_{tij}$, $\widehat{S}_Y(t)$, $\widehat{S}_T(t)$, $\widehat{M}_C(.)$ et $\widehat{M}_T(.)$.

Ces résultats ont également été repris par Blanche et al. (2013a) et adaptés afin de permettre la prise en compte de risques compétitifs dans la correction du biais de l'estimateur de l'AUC.

Méthode de Li et al.

Dans leur article, Li et al. (2018) proposent une méthode alternative pour corriger le biais de l'estimateur de l'AUC. Soient $D(t)$ la variable binaire indiquant si l'évènement est survenu avant le temps t, et $W_i = \mathbf{E}[D_i(t)]$ le risque de survenue de l'évènement d'intérêt dans un délai $[0,t]$, pour le patient i, qui servira à pondérer les patients.

Les auteurs identifient quatre scénarios possibles qui vont permettre d'attribuer un poids à chaque individu :

1. Si $Y_i > t$, on appelle le patient un « contrôle » au temps t, et le statut associé est $D_i(t) = 0$, un poids $W_i = 0$ est attribué à ce patient.

2. Si $Y_i \leq t$ et $\delta_i = 1$, on appelle le patient un « cas » au temps t, et le statut associé est $D_i(t) = 1$, un poids $W_i = 1$ est attribué à ce patient.
3. Si $Y_i = t$ et $\delta_i = 0$, on sait que $T_i > t$ et le statut du patient est $D_i(t) = 0$, le poids attribué au patient est $W_i = 0$. On notera que dans ce dernier cas, lorsque le temps est mesuré sur une échelle continue alors la probabilité théorique de ce scénario est égale à 0.
4. Si $Y_i < t$ et $\delta_i = 0$, le statut pour le patient est inconnu mais la probabilité que le patient soit un cas est $\mathrm{P}(T_i \leq t|Y_i, X_i)$ et la probabilité que le patient soit un contrôle est $\mathrm{P}(T_i > t|Y_i, X_i)$.

Ces quatre scénarios peuvent alors être résumés en une seule formule permettant de calculer le poids attribué à un individu :

$$\begin{aligned}
W_i &= \mathrm{P}(T_i \leq t|Y_i, \delta_i, X_i) \\
&= \mathbf{E}[D_i(t)|Y_i, \delta_i, X_i] \\
&= \mathbb{1}(Y_i \leq t)\delta_i + \mathbb{1}(Y_i \leq t)(1-\delta_i)\mathrm{P}(T_i \leq t|Y_i, X_i) \\
&= \mathbb{1}(Y_i \leq t)\delta_i + \mathbb{1}(Y_i \leq t)(1-\delta_i)(1-\mathrm{P}(T_i > t|Y_i, X_i)) \\
&= \mathbb{1}(Y_i \leq t)\delta_i + \mathbb{1}(Y_i \leq t)(1-\delta_i) - \mathbb{1}(Y_i \leq t)(1-\delta_i)\mathrm{P}(T_i > t|Y_i, X_i) \\
&= \mathbb{1}(Y_i \leq t) - \mathbb{1}(Y_i \leq t)(1-\delta_i)\mathrm{P}(T_i > t|Y_i, X_i) \\
&= [1-(1-\delta_i)\mathrm{P}(T_i > t|Y_i, X_i)]\mathbb{1}(Y_i \leq t)\,.
\end{aligned}$$

Il est alors possible de réécrire $\mathrm{P}(T_i > t|Y_i, X_i)$ sous la forme suivante, en utilisant les propriétés des fonctions de survie :

$$\begin{aligned}
\mathrm{P}(T_i > t|Y_i, X_i) &= \frac{\mathrm{P}(T_i > t, T_i > Y_i|X_i)}{\mathrm{P}(T_i > Y_i|X_i)} \\
&= \frac{\mathrm{P}(T_i > t|X_i)}{\mathrm{P}(T_i > Y_i|X_i)} \\
&= \frac{S_T(t|X_i)}{S_T(Y_i|X_i)}\,.
\end{aligned}$$

Ainsi

$$W_i = \left[1-(1-\delta_i)\frac{S_T(t|X_i)}{S_T(Y_i|X_i)}\right]\mathbb{1}(Y_i \leq t)\,.$$

Les auteurs proposent d'estimer les fonctions de survie conditionnelles ci-dessus à l'aide de la méthode de Kaplan-Meier pondérée par noyau :

$$\widehat{\mathrm{P}}(T_i > t|X_i) = \prod_{s\in\Omega, s\leq t}\left[1-\frac{\sum_j K_h(X_j, X_i)\mathbb{1}(Y_j = s)\delta_j}{\sum_j K_h(X_j, X_i)\mathbb{1}(Y_j \geq s)}\right],$$

où Ω est l'ensemble des Y_i distincts avec $\delta_i = 1$, et K_h est un noyau uniforme calculé sur la fenêtre d'observation h.

Il est alors possible d'estimer la sensibilité et la spécificité associées au marqueur pour un

seuil c :

$$\widehat{\mathrm{Se}}(c) = \mathrm{P}(X_i > c | T_i \leq \tau) = \frac{\sum_{i=1}^{n} \widehat{W}_i \mathbb{1}(X_i > c)}{\sum_{i=1}^{n} \widehat{W}_i},$$
$$\widehat{\mathrm{Sp}}(c) = \mathrm{P}(X_i \leq c | T_i > \tau) = \frac{\sum_{i=1}^{n} (1-\widehat{W}_i) \mathbb{1}(X_i \leq c)}{\sum_{i=1}^{n} (1-\widehat{W}_i)}.$$

L'AUC est obtenue par la formule suivante :

$$\widehat{\theta}_{Li}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \widehat{W}_i (1-\widehat{W}_j) [\mathbb{1}(X_i > X_j) + 0.5 \times \mathbb{1}(X_i = X_j)]}{\sum_{i=1}^{n} \sum_{j=1}^{n} \widehat{W}_i (1-\widehat{W}_j)}.$$

Les auteurs démontrent que cet estimateur est asymptotiquement sans biais et qu'il ne dépend pas de la taille de la fenêtre h utilisée pour calculer $K_h(.)$ (si tant est qu'aucune valeur aberrante n'est utilisée pour la taille de la fenêtre). De plus, la méthode peut être modifiée pour devenir insensible à toute transformation monotone de X en remplaçant l'approche de la fenêtre h par un « span » (proportion de patients utilisés dans l'estimation de $K_h(.)$). Il est également à noter que pour cette approche, contrairement à l'approche IPCW, tous les patients (qu'ils soient censurés ou non) ont un poids et participent au calcul de θ_{Li}, ce qui pourrait améliorer la précision de la méthode en comparaison de l'approche par IPCW. Enfin cette méthode a également l'avantage d'être totalement non paramétrique.

Concernant la construction d'un intervalle de confiance, les auteurs n'ont pas proposé une expression analytique de la variance asymptotique de cet estimateur, ils proposent d'utiliser un bootstrap de type BCa (pour « Bias-corrected and accelerated ») (Efron and Tibshirani, 1993; Davison and Hinkley, 1997), cependant étant donné que les auteurs ont démontré que leur estimateur était asymptotiquement sans biais, un bootstrap percentile plus classique peut également être utilisé.

Etude de simulation évaluant les propriétés de l'estimateur de Δ_θ en présence de censures

Design des simulations

Plusieurs études de simulation ont été conduites afin d'évaluer les performances de l'estimateur de Δ_θ en présence de censures indépendantes des valeurs du marqueur étudié. L'estimateur naïf de Δ_θ a été évalué, ainsi que la méthode de correction par IPCW (Hung and Chiang, 2010; Blanche et al., 2013a), et celle proposée par Li et al. (2018).

Pour la méthode de correction par IPCW, la formule de la variance asymptotique de l'estimateur a été utilisée pour construire les intervalles de confiance de Δ_θ pour chaque jeu de données simulé. Pour la méthode de Li et al. (2018), un boostrap percentile a été utilisé pour construire l'intervalle de confiance de Δ_θ avec 1000 réplications bootstrap pour chaque jeu de données simulé.

Ces études de simulations ont été définies en faisant varier la valeur théorique de Δ_θ (0.3 et 0.1), la proportion théorique de censures (15 %, 30 % et 60 %), ainsi que la taille de l'échantillon N (400, 1000 et 2000). Pour chacun des scénarios, le risque moyen de survenue d'évènement

dans chaque bras était égal à 0.3. Contrairement aux simulations effectuées dans l'article, les valeurs des marqueurs étaient simulées globalement pour les deux bras de traitement, puis le statut des individus au temps t déterminé en fonction des valeurs du marqueur.

Pour chaque simulation, les valeurs de marqueur étaient générées dans les deux bras à partir d'une loi normale centrée réduite $\mathcal{N}(0,1)$, les temps jusqu'à évènement et jusqu'à la censure étaient générés à partir de modèles à taux proportionnels avec un taux de base reposant sur une fonction de Weibull de telle sorte que $\lambda_T(t|X) = \frac{\beta t^{\beta-1}}{\eta^\beta}\exp(\alpha X)$ et $\lambda_C(t|X) = \frac{\nu t^{\nu-1}}{\theta^\nu}\exp(\gamma X)$.

Les différentes valeurs de paramètres de ces modèles ont été choisies de telle sorte d'obtenir les différentes configurations qui définissent les scénarios des simulations.

Les valeurs théoriques de Δ_θ étaient calculées par intégration numérique en s'appuyant sur les modèles définis précédemment.

Au total, 5000 jeux de données ont été simulés pour chaque scénario de simulation.

Résultat des simulations

Les résultats des simulations sont présentés dans les Tableaux 2.2 et 2.3. Dans tous les scénarios, le biais de l'estimateur naïf de Δ_θ augmentait avec la proportion de données censurées dans l'échantillon. Par exemple, avec $\Delta_\theta = 0.3$ et $N = 2000$, le biais de l'estimateur naïf était de 0.00104 avec 15 % de censures et de 0.01027 avec 60 % de censures. La probabilité de couverture de l'intervalle de confiance de Δ_θ pour l'estimateur naïf de Δ_θ était toujours proche de 95 %, excepté lorsque $\Delta_\theta = 0.3$ et que la proportion de censures dans l'échantillon était de 60 %.

Pour les estimateurs IPCW et de Li et al. (2018), le biais était comparable, toujours proche de 0 et équivalent ou inférieur au biais de l'estimateur naïf. Les probabilités de couverture des intervalles de confiance basés sur les estimateurs IPCW et de Li et al. (2018) étaient toujours proches de 95 % dans la majorité des scénarios, et étaient toujours équivalentes ou plus proches de 95 % que les probabilités de couverture de l'intervalle de confiance basé sur l'estimateur naïf.

Les amplitudes moyennes des intervalles de confiance basés sur les estimateurs IPCW et de Li et al. (2018) étaient toujours comparables, ou bien en faveur de l'estimateur de Li et al. (2018). Globalement, l'utilisation de Δ_θ pour quantifier les performances prédictives de marqueurs serait tout à fait envisageable dans les études présentant des censures, en utilisant la méthode de Li ou par IPCW pour corriger le biais éventuel.

2.6.3 Comparaison de Δ_θ entre différents marqueurs

La capacité prédictive de deux marqueurs (notés A et B) peut être comparée au travers d'un test statistique faisant comme hypothèses :

$$\mathrm{H}_0 : \Delta_\theta^A = \Delta_\theta^B,$$
$$\mathrm{H}_1 : \Delta_\theta^A \neq \Delta_\theta^B.$$

Tableau 2.2 – Résultats de l'étude de simulation pour $\Delta_\theta = 0.1$

		Naïf		IPCW			Li		
% cens.	N	Biais	CP	Biais	CP	WCI	Biais	CP	WCI
15%	400	0.00046	0.9480	-0.00039	0.9484	0.21	-0.00029	0.9462	0.21
	1000	0.00108	0.9490	0.00024	0.9482	0.13	0.00035	0.9492	0.13
	2000	0.00015	0.9474	-0.00070	0.9480	0.09	-0.00053	0.9496	0.09
30%	400	0.00189	0.9498	-0.00020	0.9480	0.23	-0.00026	0.9492	0.23
	1000	0.00180	0.9462	-0.00020	0.9465	0.15	0.00025	0.9488	0.14
	2000	0.00118	0.9498	-0.00086	0.9488	0.10	-0.00052	0.9502	0.10
60%	400	0.00612	0.9478	0.00113	0.9462	0.30	-0.00001	0.9432	0.27
	1000	0.00459	0.9516	-0.00018	0.9480	0.19	0.00090	0.9434	0.17
	2000	0.00411	0.9488	-0.00051	0.9486	0.13	0.00016	0.9542	0.12

Δ_θ : aire entre les courbes ROC ; N : taille de l'échantillon ; CP : probabilité de couverture ; WCI : amplitude moyenne de l'intervalle de confiance

Tableau 2.3 – Résultats de l'étude de simulation pour $\Delta_\theta = 0.3$

		Naïf		IPCW			Li		
% cens.	N	Biais	CP	Biais	CP	WCI	Biais	CP	WCI
15%	400	0.00204	0.9280	-0.00049	0.9360	0.21	-0.00053	0.9360	0.21
	1000	0.00227	0.9468	0.00034	0.9486	0.13	0.00040	0.9462	0.13
	2000	0.00104	0.9440	-0.00073	0.9446	0.09	-0.00053	0.9442	0.09
30%	400	0.00495	0.9304	-0.00004	0.9356	0.23	-0.00033	0.9390	0.22
	1000	0.00457	0.9436	0.00031	0.9448	0.14	0.00053	0.9466	0.14
	2000	0.00325	0.9432	-0.00073	0.9434	0.10	-0.00037	0.9502	0.10
60%	400	0.01095	0.9396	0.00055	0.9364	0.31	-0.00154	0.9330	0.28
	1000	0.01114	0.9320	0.00026	0.9444	0.19	0.00098	0.9466	0.17
	2000	0.01027	0.9380	-0.00008	0.9476	0.14	0.00051	0.9408	0.12

Δ_θ : aire entre les courbes ROC ; N : taille de l'échantillon ; CP : probabilité de couverture ; WCI : amplitude moyenne de l'intervalle de confiance

En notant $\boldsymbol{\Delta}_\theta = (\Delta_\theta^A, \Delta_\theta^B)'$, et en utilisant les résultats présentés dans l'article, il est possible de dire que $\widehat{\boldsymbol{\Delta}}_\theta \sim \mathcal{N}(\boldsymbol{\Delta}_\theta, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ étant la matrice de variance-covariance définie comme :

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}(\widehat{\Delta}_\theta^A) & \text{Cov}(\widehat{\Delta}_\theta^A, \widehat{\Delta}_\theta^B) \\ \text{Cov}(\widehat{\Delta}_\theta^A, \widehat{\Delta}_\theta^B) & \text{Var}(\widehat{\Delta}_\theta^B) \end{pmatrix}.$$

Comme les performances des deux marqueurs sont évaluées sur le même échantillon de patients, $\text{Cov}(\widehat{\Delta}_\theta^A, \widehat{\Delta}_\theta^B)$ est un terme non nul qu'il est nécessaire d'estimer. La variance de $\widehat{\Delta}_\theta^A - \widehat{\Delta}_\theta^B$ peut être calculée comme suit :

$$\begin{aligned} \text{Var}(\widehat{\Delta}_\theta^A - \widehat{\Delta}_\theta^B) &= \text{Var}(\widehat{\theta}_1^A - \widehat{\theta}_0^A - \widehat{\theta}_1^B + \widehat{\theta}_0^B) \\ &= \text{Var}(\widehat{\theta}_1^A) + \text{Var}(\widehat{\theta}_0^A) + \text{Var}(\widehat{\theta}_1^B) + \text{Var}(\widehat{\theta}_0^B) - 2\text{Cov}(\widehat{\theta}_1^A, \widehat{\theta}_1^B) - 2\text{Cov}(\widehat{\theta}_0^A, \widehat{\theta}_0^B) \\ &= \text{Var}(\widehat{\theta}_1^A - \widehat{\theta}_1^B) + \text{Var}(\widehat{\theta}_0^A - \widehat{\theta}_0^B). \end{aligned}$$

Ce résultat provient du fait que les AUCs des marqueurs pour deux traitements différents sont indépendantes, mais pas les AUCs des deux marqueurs pour un même traitement. Il est alors possible d'utiliser la méthode de DeLong et al. (1988) pour estimer $\text{Var}(\widehat{\theta}_1^A - \widehat{\theta}_1^B)$ et $\text{Var}(\widehat{\theta}_0^A - \widehat{\theta}_0^B)$ afin de calculer les covariances entre les AUCs qui ne sont pas indépendantes.

Si cette méthode de comparaison de marqueurs est mathématiquement valide, elle n'est pas toujours appropriée cliniquement. La Figure 2.15 présente deux situations. Sur le graphique de gauche est présentée la situation pour laquelle les courbes ROC d'un marqueur sont emboîtées entre les courbes ROC du deuxième marqueur. Dans ce cas précis, il est possible de comparer les performances des deux marqueurs sans tenir compte des coûts cliniques associés aux traitements étudiés car peu importe ceux-ci. En effet, quel que soit le seuil retenu pour attribuer les traitements, le marqueur ayant la plus grande valeur de Δ_θ aura de meilleures performances prédictives.

Sur le graphique de droite est présentée la situation pour laquelle les courbes ROC des deux marqueurs ne sont pas emboîtées. Dans ce cas, selon le seuil retenu des marqueurs, le marqueur à utiliser pour le choix du traitement peut varier. Le choix du seuil étant fonction, entre autres, des coûts cliniques des traitements, le test statistique présenté précédemment peut ne pas être suffisant pour dire qu'un marqueur a une utilité clinique plus importante que l'autre marqueur. Il serait donc intéressant d'adapter des approches par courbes ROC développées pour l'analyse des marqueurs diagnostiques et pronostiques (Subtil and Rabilloud, 2015) permettant de tenir compte de l'utilité clinique à l'étude des marqueurs prédictifs.

Etude de simulation évaluant les propriétés du test de Wald pour les courbes ROC emboîtées

Une étude de simulation a été réalisée afin d'évaluer les propriétés du test de Wald (quantification du risque α sous l'hypothèse nulle, puissance) dans la situation où les courbes ROC sont emboîtées (graphique de gauche de la Figure 2.15)

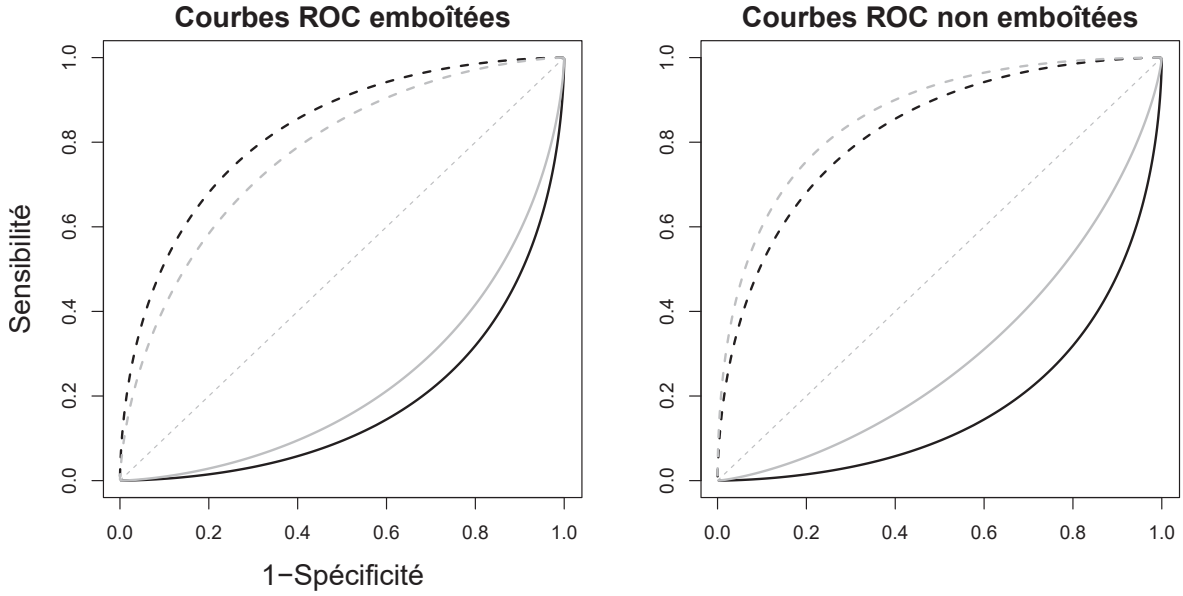


FIGURE 2.15 – Comparaison de marqueurs lorsque les courbes ROC sont emboîtées ou non Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence ; En noir : marqueur A ; En gris : marqueur B

Design des simulations

Pour évaluer les propriétés du test statistique, un scénario dans lequel le risque moyen de survenue d'évènement dans les deux bras était égal à 0.5 a été envisagé. La première étude de simulation visait à vérifier que, sous l'hypothèse nulle, le risque α était bien toujours de 5 %. Pour ce faire, plusieurs paramètres de simulation ont été envisagés :

- $\Delta_\theta^A = \Delta_\theta^B = \{0.6, 0.4, 0.2\}$;
- L'effectif total de l'essai $N = \{200, 400, 1000\}$.

La deuxième étude de simulation visait à quantifier la puissance du test en fonction des écarts entre Δ_θ^A et Δ_θ^B, pour les mêmes valeurs de N que celles considérées dans la première étude de simulation :

- Ecart de 0.2 entre Δ_θ^A et Δ_θ^B : $\{0.6$ vs. $0.4, 0.5$ vs. $0.3, 0.4$ vs. $0.2, 0.3$ vs. $0.1\}$;
- Ecart de 0.1 entre Δ_θ^A et Δ_θ^B : $\{0.6$ vs. $0.5, 0.5$ vs. $0.4, 0.4$ vs. $0.3, 0.3$ vs. $0.2, 0.2$ vs. $0.1\}$.

Chaque combinaison de paramètres a été évaluée sur 10 000 jeux de données. Pour chacun de ces jeux de données les valeurs de marqueur étaient simulées à partir de lois normales.

Résultats des simulations

Le Tableau 2.4 présente les résultats de la première étude de simulation. Peu importe l'effectif ou bien les valeurs de Δ_θ^A et Δ_θ^B, le risque α est proche de 5 % dans tous les cas de figure sous l'hypothèse nulle.

Les Tableaux 2.5 et 2.6 présentent les résultats de la deuxième étude de simulation permettant de quantifier la puissance selon plusieurs combinaisons de paramètres. Dans tous les cas de figure, la puissance décroît avec l'effectif total de l'essai, ainsi qu'avec la réduction de l'écart

Tableau 2.4 – Résultats de la première étude de simulation

	N	α
$\Delta_\theta^A = \Delta_\theta^B = 0.6$	200	0.0482
	400	0.0505
	1000	0.0486
$\Delta_\theta^A = \Delta_\theta^B = 0.4$	200	0.0505
	400	0.0496
	1000	0.0544
$\Delta_\theta^A = \Delta_\theta^B = 0.2$	200	0.0507
	400	0.0489
	1000	0.0524

Δ_θ : aire entre les courbes ROC ; N : taille de l'échantillon

entre Δ_θ^A et Δ_θ^B. Par exemple, lorsqu'un écart de 0.2 était attendu entre Δ_θ^A et Δ_θ^B, et que $\Delta_\theta^A = 0.6$ et $\Delta_\theta^B = 0.4$ alors un effectif de 400 était suffisant pour obtenir une puissance supérieure à 80 %. En revanche, dans les autres cas de figure où $\Delta_\theta^A - \Delta_\theta^B = 0.2$, il fallait plus de 400 patients au total dans l'essai pour obtenir une puissance supérieure à 80 %.

En réduisant l'écart entre Δ_θ^A et Δ_θ^B à 0.1, un effectif de 1000 n'était jamais suffisant pour avoir une puissance supérieure à 80 %. Ceci souligne à nouveau la nécessité d'effectifs élevés pour l'évaluation et la comparaison de marqueurs prédictifs.

Tableau 2.5 – Résultats de la deuxième étude de simulation ($\Delta_\theta^A - \Delta_\theta^B = 0.2$)

	N	Puissance
$\Delta_\theta^A = 0.6$ vs. $\Delta_\theta^B = 0.4$	200	0.5316
	400	0.8304
	1000	0.9950
$\Delta_\theta^A = 0.5$ vs. $\Delta_\theta^B = 0.3$	200	0.4931
	400	0.7874
	1000	0.9924
$\Delta_\theta^A = 0.4$ vs. $\Delta_\theta^B = 0.2$	200	0.4518
	400	0.7356
	1000	0.9850
$\Delta_\theta^A = 0.3$ vs. $\Delta_\theta^B = 0.1$	200	0.4012
	400	0.7090
	1000	0.9767

Δ_θ : aire entre les courbes ROC ; N : taille de l'échantillon

2.7 Bilan du chapitre 2

L'objectif de ce chapitre était de définir ce qu'était un marqueur prédictif et de présenter des méthodes permettant de quantifier la capacité prédictive d'un marqueur. Il a été vu que trois grandes approches se distinguaient. La première d'entre elles repose sur la construction d'un modèle permettant de prédire le risque de survenue de l'évènement d'intérêt incluant une interaction marqueur-traitement. Un test statistique sur le coefficient d'interaction permet alors

Tableau 2.6 – Résultats de la deuxième étude de simulation ($\Delta_\theta^A - \Delta_\theta^B = 0.1$)

	N	Puissance
$\Delta_\theta^A = 0.6$ vs. $\Delta_\theta^B = 0.5$	200	0.1832
	400	0.3262
	1000	0.6688
$\Delta_\theta^A = 0.5$ vs. $\Delta_\theta^B = 0.4$	200	0.1707
	400	0.2914
	1000	0.6047
$\Delta_\theta^A = 0.4$ vs. $\Delta_\theta^B = 0.3$	200	0.1604
	400	0.2691
	1000	0.5589
$\Delta_\theta^A = 0.3$ vs. $\Delta_\theta^B = 0.2$	200	0.1449
	400	0.2480
	1000	0.5189
$\Delta_\theta^A = 0.2$ vs. $\Delta_\theta^B = 0.1$	200	0.1076
	400	0.2293
	1000	0.5044

Δ_θ : aire entre les courbes ROC ; N : taille de l'échantillon

de détecter des marqueurs pour lesquels la différence de risques varie en fonction des valeurs du marqueur, et donc de détecter des marqueurs potentiellement prédictifs. Cependant cette approche souffre de plusieurs inconvénients. Il a été notamment vu que, selon la structure retenue pour le modèle (additive ou multiplicative), ou bien selon l'échelle du marqueur, le coefficient d'interaction n'avait plus la même interprétation et pouvait mener à des confusions. Par ailleurs, ces méthodes nécessitent de modéliser correctement l'évolution du risque selon les valeurs de marqueurs, et l'interaction entre le marqueur et le traitement.

La deuxième approche est basée sur l'évaluation du bénéfice individuel des patients, dans le but de construire une courbe ROC permettant de discriminer les patients bénéficiant du traitement innovant, de ceux bénéficiant du traitement de référence. Bien que très attrayantes en matière d'interprétation (on mesure directement la capacité du marqueur à identifier les patients qui retirent un bénéfice du traitement innovant), ces méthodes reposent sur des hypothèses parfois difficilement justifiables, et surtout très complexes rendant leur utilisation en pratique très limitée (Janes et al., 2015b).

Enfin la troisième approche n'est pas basée sur l'évaluation du bénéfice individuel des patients. L'objectif est de quantifier le caractère prédictif global du marqueur en recherchant des marqueurs pour lesquels la différence de risques moyenne entre les deux bras de traitement varie selon les valeurs du marqueur. Cette approche rend les indicateurs moins intuitifs en matière d'interprétation, cependant elle est valide dans le cadre des essais cliniques randomisés. C'est d'ailleurs dans le cadre de cette approche qu'un indicateur non paramétrique a été proposé, reposant sur un extension de l'utilisation des courbes ROC classiques à la détection des marqueurs prédictifs. Cet indicateur est utile dans les premières étapes d'identification et d'évaluation d'un marqueur prédictif. La validité de cette approche n'est pour l'instant démontrée que lorsque les risques moyens d'évènement dans les deux bras sont égaux.

La quantification du caractère prédictif global d'un marqueur n'est pas suffisante pour valider son utilisation dans la pratique clinique. Lorsque le marqueur est quantitatif, il est nécessaire de déterminer une règle de décision basée sur la valeur du marqueur afin d'allouer chaque traitement seulement aux patients qui en retirent le plus grand bénéfice. La détermination de ce seuil de marqueur doit tenir compte de l'efficacité de ces traitements, des risques moyens de survenue d'évènement dans chaque bras de traitement, mais également de leurs toxicités et donc de leur impact sur la qualité de vie des patients. Ces notions sont abordées dans le chapitre suivant, et des méthodes permettant d'estimer ce seuil sont présentées.

Chapitre 3

Seuil optimal d'un marqueur prédictif

Une fois qu'un biomarqueur prédictif quantitatif a été identifié, la question de l'estimation d'une valeur seuil du marqueur au-delà duquel un traitement sera préféré se pose.

Pour ce faire, il est nécessaire de définir une fonction reflétant l'état de santé dans la population cible, et reposant sur une règle de décision régissant l'administration des traitements comparés. Par exemple, une règle de décision pourrait être : « Je donne le traitement de référence aux patients dont la valeur de marqueur est supérieure à un seuil c, et je donne le traitement innovant aux autres ». Cette fonction, qui dépend d'une valeur seuil c, devra alors être maximisée afin d'identifier ce que l'on appellera le seuil optimal du marqueur.

A ce jour, très peu de méthodes ont été proposées pour estimer ce seuil ainsi que son intervalle de confiance. Pourtant, la définition de ce seuil est essentielle pour la prise de décision médicale.

3.1 Proposition d'une approche pour l'estimation du seuil optimal

La méthode décrite ci-après repose sur l'expression d'une fonction d'utilité qui quantifie l'utilité moyenne dans la population cible lorsqu'une stratégie d'allocation de traitement basée sur un marqueur est utilisée, en tenant compte à la fois de l'efficacité et des toxicités des différentes options de traitement. Par exemple, l'utilité moyenne d'une population peut être mesurée par l'espérance de vie des patients dans une situation donnée, l'espérance de vie pondérée par la qualité de vie, ou bien la satisfaction moyenne des patients pour un état de santé donné (Sox et al., 2013).

Ce qui motive l'analyse d'utilité est l'hypothèse que les décideurs choisissent la règle d'allocation de traitement dans le but de maximiser l'utilité moyenne d'une population. Par exemple, quand l'utilité est mesurée par l'espérance de vie, alors l'analyse d'utilité identifie la stratégie de traitement qui maximise l'espérance de vie d'une population (Sox et al., 2013). Quand une stratégie basée sur un marqueur est utilisée pour guider le choix du traitement, le seuil optimal du marqueur est la valeur de marqueur qui maximise l'utilité moyenne dans la population-cible ; il s'agit donc de la valeur de marqueur qui maximise la fonction d'utilité.

3.1.1 Définition de la fonction d'utilité

On se place dans le contexte simple où l'objectif est de choisir entre le traitement de référence ($Z = 0$) et le traitement innovant ($Z = 1$). Pour les développements qui suivront, nous ferons l'hypothèse que le traitement innovant est recommandé pour des valeurs de marqueur inférieures au seuil de marqueur c, et que le traitement de référence est recommandé pour des valeurs de marqueur supérieures à c.

Sous ces hypothèses, il est possible de définir quatre groupes de patients selon s'ils ont ou non développé l'évènement d'intérêt et selon le traitement qu'ils ont reçu :

- Population $0Y$: Patients qui ont développé l'évènement dans le bras référence ;
- Population $0\bar{Y}$: Patients qui n'ont pas développé l'évènement dans le bras référence ;
- Population $1Y$: Patients qui ont développé l'évènement dans le bras innovant ;
- Population $1\bar{Y}$: Patients qui n'ont pas développé l'évènement dans le bras innovant.

On définit $U_{zl}(x)$ comme l'utilité moyenne associée aux patients dans la population zl et ayant une valeur de marqueur égale à x, avec $z = 0,1$ pour les patients traités avec le traitement de référence ou innovant respectivement, et $l = \bar{Y}, Y$ pour les patients qui ne développent pas l'évènement ou bien qui le développent respectivement. Ces mesures d'utilité servent à quantifier la préférence pour un état de santé donné. Le Tableau 3.1 permet de définir les utilités moyennes en fonction du traitement recommandé (Z), du critère de jugement observé (Y), ainsi que du marqueur (X).

Tableau 3.1 – Présentation des utilités moyennes pour les quatre groupes de patients

	$Z = 0$ $(X > c)$	$Z = 1$ $(X \leq c)$
$Y = 0$	$U_{0\bar{Y}}(X)$	$U_{1\bar{Y}}(X)$
$Y = 1$	$U_{0Y}(X)$	$U_{1Y}(X)$

La fonction $U(c)$ qui quantifie l'utilité moyenne dans la population cible, sachant la règle de décision définie précédemment, s'exprime comme suit :

$$\begin{aligned} U(c) = &\int_{c}^{+\infty} \mathrm{P}(Y=1, X=x|Z=0) \times U_{0Y}(x) + \mathrm{P}(Y=0, X=x|Z=0) \times U_{0\bar{Y}}(x)\, dx \\ &+ \int_{-\infty}^{c} \mathrm{P}(Y=1, X=x|Z=1) \times U_{1Y}(x) + \mathrm{P}(Y=0, X=x|Z=1) \times U_{1\bar{Y}}(x)\, dx, \end{aligned} \tag{3.1}$$

où la première intégrale évalue l'utilité moyenne des patients recevant le traitement de référence en fonction de la probabilité qu'ils développent ou non l'évènement d'intérêt et des utilités associées à ces états. La seconde intégrale effectue le même calcul pour les patients recevant le traitement innovant.

Les différentes utilités définies dans cette équation sont souvent exprimées sous la forme de coûts dans la littérature (Vickers et al., 2007; Janes et al., 2014b; Huang et al., 2015). Alors que les utilités mesurent une quantité préférable, les coûts mesurent une quantité non souhaitable

pour le patient, ils ne sont donc pas exprimés sur la même échelle. Le Tableau 3.2 adapte le Tableau 3.1 au cas du coût moyen d'une stratégie de traitement en utilisant les notations données dans l'article de Janes et al. (2014b).

Tableau 3.2 – Présentation des coûts moyens pour les quatre groupes de patients

	$Z=0$ $(X>c)$	$Z=1$ $(X\leq c)$
$Y=0$	0	$C_Z(X)$
$Y=1$	$C_Y(X)$	$C_Z(X)+C_Y(X)+C_{++}(X)$

Il est considéré arbitrairement que la situation « absence d'événement » et « utilisation du traitement de référence » est une situation à coût nul. Les autres termes de coûts correspondent aux surcoûts liés à l'utilisation du traitement innovant, au développement de l'événement, ou à l'interaction entre ces deux surcoûts. Comme minimiser le coût moyen revient à maximiser l'utilité moyenne, il est possible de définir $C_Y(X) = -[U_{0Y}(X) - U_{0\bar{Y}}(X)]$ comme le coût d'un évènement dans le bras de référence sachant X, $C_Y(X) + C_{++}(X) = -[U_{1Y}(X) - U_{1\bar{Y}}(X)]$ comme le coût d'un évènement dans le bras innovant sachant X, et $C_Z(X) = -[U_{1\bar{Y}}(X) - U_{0\bar{Y}}(X)]$ comme le coût supplémentaire du traitement innovant par rapport au traitement de référence en l'absence d'évènement sachant X.

Dans l'expression (3.1), chacune des utilités peut dépendre du marqueur étudié X. Cela pourrait se justifier par exemple si le marqueur étudié est l'âge, il y a alors de grandes chances que l'impact de certaines toxicités ne soit pas le même pour des patients jeunes ou bien pour des patients âgés. Il est possible de simplifier l'expression de $U(.)$ en imposant des contraintes sur les différentes utilités (ou coûts). Par exemple, Vickers et al. (2007) font les hypothèses suivantes :

$$\begin{aligned}
C_Y(X) = C_Y &\Leftrightarrow -[U_{0Y}(X) - U_{0\bar{Y}}(X)] = -[U_{0Y} - U_{0\bar{Y}}]\,,\\
C_Z(X) = C_Z &\Leftrightarrow -[U_{1\bar{Y}}(X) - U_{0\bar{Y}}(X)] = -[U_{1\bar{Y}} - U_{0\bar{Y}}]\,,\\
C_{++}(X) = C_{++} = 0 &\Leftrightarrow -[U_{1Y}(X) - U_{1\bar{Y}}(X)] = -[U_{1Y} - U_{1\bar{Y}}] = -[U_{0Y} - U_{0\bar{Y}}]\,,
\end{aligned}$$

où les coûts et utilités ne dépendent plus du marqueur étudié, et où la troisième ligne signifie que le coût d'un évènement est le même dans les deux bras de traitement. Cette dernière hypothèse est assez classique, bien qu'elle ne s'applique pas dans certains cas spécifiques. Par exemple, dans le cas du traitement préventif du cancer du sein, l'utilisation du tamoxifène permet de réduire la probabilité de survenue d'un cancer du sein. Cependant, l'utilisation de tamoxifène peut n'être efficace que pour prévenir les cancers du sein à récepteurs d'œstrogène positifs (ER+), ce qui fait que les cancers du sein diagnostiqués sous tamoxifène ont plus de chance d'être des cancers ER- qui sont généralement plus difficiles à traiter ; ainsi le coût de l'évènement « développer un cancer » est plus important pour les patientes traitées par tamoxifène que pour celles qui ne reçoivent pas ce traitement (Janes et al., 2014b).

Sous les hypothèses de Vickers et al. (2007), la fonction d'utilité (3.1) peut s'écrire sous la

forme suivante :

$$U(c) \propto F(c)^{(0Y)} \times \rho_0 - F(c)^{(1Y)} \times \rho_1 - F(c) \times \frac{C_Z}{C_Y},$$

où $F(.)$ ets la fonction de répartition marginale du marqueur et $F(.)^{(zl)}$ est la fonction de répartition du marqueur pour le groupe de patients zl. L'intérêt ici est que les quatre utilités définies dans le Tableau 3.1 sont résumées sous la forme du ratio de coûts $\frac{C_Z}{C_Y}$. Ce ratio a une interprétation intéressante, d'après Vickers et al. (2007) il s'agit de la différence de risque d'évènement entre les deux traitements à partir de laquelle il est acceptable de traiter un patient avec le traitement le plus toxique. L'inverse de ce ratio peut également être interprété comme le nombre de patients qu'un clinicien est prêt à traiter avec le traitement le plus toxique pour éviter un évènement de plus en comparaison de la stratégie de traitement la moins toxique.

3.1.2 Article accepté dans la revue Statistical Methods in Medical Research

Cet article a été accepté par la revue Statistical Methods in Medical Research et présente de manière approfondie la fonction d'utilité définie précédemment, détaille sa justification, propose une méthode d'inférence associée au seuil optimal, et en évalue les performances. Les informations supplémentaires disponibles sur le site de la revue sont présentées en Annexe C.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–15

DOI: 10.1177/0962280218821394
journals.sagepub.com/home/smm
SAGE

A Bayesian method to estimate the optimal threshold of a marker used to select patients' treatment

Yoann Blangero,[1,2] Muriel Rabilloud,[1,2] René Ecochard[1,2] and Fabien Subtil[1,2]

Abstract

The use of a quantitative treatment selection marker to choose between two treatment options requires the estimate of an optimal threshold above which one of these two treatments is preferred. Herein, the optimal threshold expression is based on the definition of a utility function which aims to quantify the expected utility of the population (e.g. life expectancy, quality of life) by taking into account both efficacy (success or failure) and toxicity of each treatment option. Therefore, the optimal threshold is the marker value that maximizes the expected utility of the population. A method modelling the marker distribution in patient subgroups defined by the received treatment and the outcome is proposed to calculate the parameters of the utility function so as to estimate the optimal threshold and its 95% credible interval using the Bayesian inference. The simulation study found that the method had low bias and coverage probability close to 95% in multiple settings, but also the need of large sample size to estimate the optimal threshold in some settings. The method is then applied to the PETACC-8 trial that compares the efficacy of chemotherapy with a combined chemotherapy + anti-epidermal growth factor receptor in stage III colorectal cancer.

Keywords

Biomarker, threshold, expected utility, treatment selection, Bayesian, predictive marker

1 Introduction

Biomarkers help improve patient outcomes by targeting treating those who are likely to benefit from a given treatment and avoiding treating those who would not have great interest in precision medicine (e.g. oncology, cardiology). "Treatment selection" biomarkers, often called "predictive" biomarkers,[1,2] are generally sought for when the benefit of a new treatment in comparison to a reference treatment is considered, and that this benefit is suspected to vary according to the characteristics of the patients. One example of a binary treatment selection biomarker is the absence of KRAS gene mutation that predicts the benefit of a combined chemotherapy + epidermal growth factor receptor (EGFR) inhibitor over chemotherapy alone in metastatic colorectal cancer; patients with tumors harboring mutated KRAS exon 2 are known to be resistant to EGFR inhibitors, whereas patients with KRAS wild-type tumors do benefit from the combined treatment.[3]

While it is easy to identify new binary treatment selection biomarkers, this is more difficult for quantitative ones. Various methods have been proposed to assess quantitative treatment selection biomarkers by modelling the risk of event given the treatment options and the biomarker values,[4–7] or by modelling the benefit to use one treatment rather than the other one (i.e. the difference in outcomes) for a patient given its biomarker value.[8,9] Classically, modelling the risk is performed using logistic regression models or survival regression models (e.g. Cox models), by introducing an interaction between the biomarker and the treatments. Modelling the benefit is a harder task. Indeed, the individual difference in outcomes is unobservable because each patient can receive only one treatment, that is

[1]Service de Biostatistique-Bioinformatique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France
[2]Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, Villeurbanne, France

Corresponding author:
Yoann Blangero, Service de Biostatistique-Bioinformatique, Pôle Santé Publique, Hospices Civils de Lyon, 162 avenue Lacassagne, F-69003, Lyon, France.
Email: yoann.blangero@chu-lyon.fr

why strong assumptions have to be made in order to estimate this individual benefit. For example, Huang et al.[8] made the monotonicity assumption (one treatment is always at least as efficient as the other one) to estimate the individual benefit. Zhang et al.[9] relaxed the latter assumption by assuming that the potential outcomes are independent given observed covariates which means that the benefit for a patient can be calculated by comparing its outcome to the outcome of patients that are similar given the observed covariates, and that these observed covariates are sufficient to explain the dependence between the potential outcomes.

After identification of a quantitative treatment selection biomarker, its use in clinical practice needs the determination of a threshold to help clinicians decide which treatment to administrate to a given patient.[10] Some of the above-cited papers define a threshold for the difference in risks (between the two treatment options) above or below which one of the two treatment options should be preferred.[5–7] In order for this threshold to be optimal (or clinically relevant), some of these methods were extended to take into account the relative toxicities of both treatment options[5,11] as well as the impact of treatment failure. The conversion of the difference in risks threshold to a threshold on the marker scale is straightforward but relies on a good calibration of the model used to predict the risk, or the difference in risks,[12,13] and on modelling assumptions that may be difficult to check (e.g. the increase of the biomarker value must lead to a linear increase of the logit of the risk in a logistic regression). To the best of our knowledge, no method of inference has been proposed in the literature for estimating a threshold directly on the marker scale.

These remarks have motivated the development of a new method to estimate the optimal threshold of a treatment selection biomarker, based on the maximization of a utility function and on the modelling of the marker distribution rather than the risk of event. A Bayesian inference method was used in order to obtain a punctual estimation of the threshold and its credible interval. In our opinion, modelling the marker distribution has the advantage that multiple tools exist to check the adequacy of a theoretical distribution fitted on the observed marker distribution,[14–16] contrary to risk models for which it may be difficult to check and correct calibration issues. Moreover, an extension of decision curves,[17] initially proposed for the analysis of diagnostic and prognostic markers, is presented and its connection with a previous approach[18] is discussed.

A simulation study was conducted to evaluate the bias in the optimal threshold estimate, the coverage probability and the mean width of its credible interval. The method was then applied to estimate the threshold of the DDR2 gene expression level when this marker is used to choose between FOLFOX4 and FOLFOX4 + cetuximab in stage III colon adenocarcinoma.

2 Method

The method described hereafter relies on the expression of a utility function that quantifies the expected utility of the target population when using a marker-based treatment strategy by taking into account both the efficacy and toxicity of the treatment options. The motivation behind utility analysis is the assumption that decision-makers choose the most preferable treatment option for each patient in order to maximize the expected utility of the population. For example, when the utility is measured by the length of the patient's life, the expected utility analysis identifies the most preferable treatment-strategy to maximize the life expectancy of the population.[10] When a marker-based treatment strategy is used to guide treatment decision, the marker optimal threshold is the marker value that maximizes the expected utility of the target population; it is therefore the marker value that maximizes the utility function.

2.1 Expected utility function

Let us consider the simple context in which the task is to decide between two treatment options, referred to as "innovative treatment" ($T=1$) and "reference treatment" ($T=0$). The binary event of interest is denoted by E, where $E=1$ indicates the presence of the event of interest in a post-treatment interval, and $E=0$ (or $\bar{E}$) indicates its absence. For example, E might be an indicator of cancer recurrence or death. Let $\rho_0 = P(E=1|T=0)$ and $\rho_1 = P(E=1|T=1)$, indicating the mean risk of event under each treatment. The mean risk of event measures the efficacy of each treatment option. At equal efficacy, it is assumed that the harm of the innovative treatment is greater than that of the reference treatment; if the harm of the innovative treatment is lower than that of the reference treatment the following developments could be adapted (see online Supplemental Material). The candidate marker, denoted by X, is a quantitative measurement, or it may be a score that combines multiple measurements. It is assumed that the innovative treatment is recommended for marker values lower than the marker threshold c, and the reference treatment is recommended for marker values greater than the

marker threshold. If the innovative treatment is recommended for marker values greater than the marker threshold c, the following developments could also be adapted (see online Supplemental Material).

Under these assumptions, four populations are defined:

- Population $0E$: Patients who developed the event of interest under the reference treatment ($X > c$)
- Population $0\bar{E}$: Patients who did not develop the event of interest under the reference treatment ($X > c$)
- Population $1E$: Patients who developed the event of interest under the innovative treatment ($X \leq c$)
- Population $1\bar{E}$: Patients who did not develop the event of interest under the innovative treatment ($X \leq c$).

Let $U_{ij}(x)$ define the utility associated with a patient in the population ij and having a marker value equal to x with $i = 0, 1$ for patients treated with the reference or the innovative treatment, respectively, and $j = E, \bar{E}$ for patients who developed the event of interest or who did not develop the event, respectively.

The expected utility function, $U(c)$, that quantifies the mean utility given the above-mentioned treatment decision rule for threshold c, is expressed as

$$\begin{aligned} U(c) = &\int_c^{+\infty} P(E=1, X=x|T=0) \times U_{0E}(x) + P(E=0, X=x|T=0) \times U_{0\bar{E}}(x)\, dx \\ &+ \int_{-\infty}^{c} P(E=1, X=x|T=1) \times U_{1E}(x) + P(E=0, X=x|T=1) \times U_{1\bar{E}}(x)\, dx \end{aligned}$$

This function quantifies the expected utility that corresponds to a weighted sum of utilities for a given threshold c. As the marker is measured before any treatment intervention, the marker distribution and the treatments are independent, therefore it is assumed that $P(X \leq c|T=0) = P(X \leq c|T=1) = P(X \leq c) = F_X(c), \quad \forall c$.

Then the utility function can be expressed as

$$\begin{aligned} U(c) = &\int_c^{+\infty} f_X(x)[\rho_0(x) \times U_{0E}(x) + \{1-\rho_0(x)\} \times U_{0\bar{E}}(x)]\, dx \\ &+ \int_{-\infty}^{c} f_X(x)[\rho_1(x) \times U_{1E}(x) + \{1-\rho_1(x)\} \times U_{1\bar{E}}(x)]\, dx \end{aligned} \tag{1}$$

where $f_X(x)$ is the probability density function of the marker X, $\rho_0(x) = P(E=1|X=x, T=0)$ is the risk of event under the reference treatment given the marker value $X=x$, and $\rho_1(x) = P(E=1|X=x, T=1)$ is the risk of event under the innovative treatment given the marker value $X=x$.

It is assumed, with very little restriction, that $U_{0\bar{E}}(x) > U_{0E}(x)$ and $U_{1\bar{E}}(x) > U_{1E}(x)$, which means that this is preferable not to develop the event of interest in either treatment arm. Here, the utilities are allowed to depend on the marker value x. For example, age could be a candidate marker in some settings, and it is likely that utilities will differ between young patients and older ones.

With a little algebra manipulation, one can see that cancelling the derivative (with respect to c) of the expected utility function leads to the following equation

$$-f_X(c)[\rho_0(c)U_{0E}(c) + \{1-\rho_0(c)\}U_{0\bar{E}}(c)] + f_X(c)[\rho_1(c)U_{1E}(c) + \{1-\rho_1(c)\}U_{1\bar{E}}(c)] = 0$$

So, the expected utility is maximal for the threshold value that verifies the following equality[11]

$$\rho_0(c)[U_{0E}(c) - U_{0\bar{E}}(c)] - \rho_1(c)[U_{1E}(c) - U_{1\bar{E}}(c)] = U_{1\bar{E}}(c) - U_{0\bar{E}}(c) \tag{2}$$

This expression helps define the optimal difference in risks threshold from which one of the two treatment options is preferred.

In practice, it may be complex to specify meaningful quantities for each utility. Vickers et al.[5] made some assumptions on utilities to avoid having to estimate the four utilities separately. They assume that $U_{0E}(x) = U_{0E}$, $U_{0\bar{E}}(x) = U_{0\bar{E}} = 1$, $U_{1E}(x) = U_{1E}$ and $U_{1\bar{E}}(x) = U_{1\bar{E}}$, which means that the utilities are not allowed to depend on the marker value, and that each utility is expressed relatively to the others. They also assume that $U_{1E} - U_{1\bar{E}} = U_{0E} - U_{0\bar{E}}$, which means that the cost associated with the development of an event is the same in either arm.

Under these assumptions, it can be shown (details in online Supplemental Material) that the utility function (1) is proportional to

$$U(c,r) \propto F_X(c)[P(E=1|X \leq c, T=0) - P(E=1|X \leq c, T=1) - r] + A \quad (3)$$

In this expression, $r = \frac{treatment}{event}$, $treatment = U_{1\bar{E}} - U_{0\bar{E}} = U_{1E} - U_{0E}$ denotes the additional cost of the innovative treatment regarding the reference one, $event = U_{0E} - U_{0\bar{E}} = U_{1E} - U_{1\bar{E}}$ denotes the cost associated with the development of the event of interest in either arm and $A = -\frac{U_{0E}\rho_0 + U_{0\bar{E}}(1-\rho_0)}{U_{0E} - U_{0\bar{E}}}$ is an additive constant relative to c. As one can see, applying the Vickers et al.[5] assumptions directly to equation (2), or maximizing the utility function presented in equation (3), leads to the expression of the optimal difference in risks threshold as defined by Vickers et al. in their original paper[5] (details in online Supplemental Material)

$$\rho_0(c) - \rho_1(c) = r \quad (4)$$

This equation means that a patient should be treated with the reference treatment if the difference in risks of event between the two treatment arms exceeds the ratio r. This result is interesting as it gives a simple interpretation of the ratio r which helps to define the utilities. This is the difference in risks of event from which it is acceptable to use the most harmful treatment option. The inverse of the ratio has also a meaning; this corresponds to the number of patients that a clinician is ready to treat with the most harmful treatment option to avoid one additional event-case compared with the less harmful treatment option.

As these assumptions simplify the utility function and that r is meaningful, the same assumptions are made in the following sections. An optimal threshold can be estimated for each r value; however, defining the treatment allocation using the marker is only relevant for a range of r values. This range can be defined using decision curves.

2.2 Decision curves

Decision curves were originally designed for diagnostic or prognostic markers.[17] Herein, it helps to compare the expected utility of using the marker to guide the treatment choice with the expected utility of extreme strategies (''Treat everyone with the innovative treatment'': $U_{T=1}(r)$ or ''Treat everyone with the reference treatment'': $U_{T=0}(r)$) for multiple r settings.

Based on equation (3), it is possible to define the expected utility of extreme strategies as

$$U_{T=1}(r) = \lim_{c \to +\infty} U(c,r) \propto \rho_0 - \rho_1 - r + A$$
$$U_{T=0}(r) = \lim_{c \to -\infty} U(c,r) \propto A$$

In fact, decision curves enable a sensitivity analysis to be conducted using the r ratio as the sensitivity parameter. The relevant range of r values is defined when the marker-based strategy is better than any of the two extreme strategies.

An extension of decision curves was proposed by Baker et al.[19] for diagnostic markers and is named ''relative utility'' curves. Huang et al.[18] proposed to apply this concept to treatment selection markers. In this case, these relative utility curves standardize the utility of the marker-based strategy in order to give to its value a meaningful interpretation (0: useless marker, 1: perfect marker). The expression of the relative utility function is proposed in the online Supplemental Material.

2.3 Estimation method

In this section, it is assumed that the marker is measured before treatment administration in a controlled parallel randomized clinical trial with two treatment arms, and that the treatment allocation does not depend on the marker values. Such a study ensures that the treatment efficacies are measured without confounders, and that $P(X \leq c|T=0) = P(X \leq c|T=1) = P(X \leq c)$. The latter formula is called the ''randomization constraint'', and we demonstrate later that this constraint is useful in the estimation method.

Previously reported methods propose to estimate the optimal difference in risks threshold by modelling the risk of event given the marker values and the treatment arms.[5,7,11] The optimal marker threshold can be derived from a risk model by searching the marker value that corresponds to the optimal difference in risks threshold, but it relies

on the good calibration of the model (which is not always easy to control), and no inference method was proposed to obtain a confidence interval of the marker threshold.

Herein, an alternative approach that does not require calibration is proposed; it is based on modelling the marker distribution in each group of patients previously defined rather than the risk of event given the treatments and the marker values.

Simply applying Bayes' theorem, and with a little algebra manipulation, the utility function (3) can be expressed as

$$\begin{aligned} U(c,r) \propto F_{X_{0E}}(c)\rho_0 - F_{X_{1E}}(c)\rho_1 - r \times [P(T=0)\{F_{X_{0E}}\rho_0 + F_{X_{0\bar{E}}}(1-\rho_0)\} \\ + P(T=1)\{F_{X_{1E}}\rho_1 + F_{X_{1\bar{E}}}(1-\rho_1)\}] + A \end{aligned}$$

where $F_{X_{ij}}(c)$ is the cumulative distribution function of the marker in the population ij. $P(T=0)$ and $P(T=1)$ are fixed by the study design. The quantities ρ_0 and ρ_1 are estimated by the mean risk of event in each arm. Empirical cumulative distribution functions may be used to estimate the marker cumulative distribution functions $F_{X_{0E}}(c)$, $F_{X_{0\bar{E}}}(c)$, $F_{X_{1E}}(c)$ and $F_{X_{1\bar{E}}}(c)$; however, depending on the sample size, it could lead to an important uncertainty in the optimal threshold estimate, and the only way to obtain a confidence interval would be by applying bootstrap methods. Bootstrap methods[20] have been proposed to obtain the confidence interval of the optimal threshold of a diagnostic marker; however, it was shown by Schisterman and Perkins[21] that it may have a poor coverage performance in some situations. This is why modelling the cumulative distribution functions of the marker has been preferred.[22,23] The proposed approach relies on finding, for each of the four aforementioned populations, a parametric distribution that fits well the data.

In practice, there is no need to fit each of the four distributions. Using the randomization constraint, one of the four marker cumulative distribution functions can be defined indirectly by the three others and the mean risks of event. We chose to express the most difficult distribution to fit, with respect to the three others. For example, if $F_{X_{0E}}(c)$ is the most difficult distribution to estimate, the randomization constraint states that

$$F_{X_{0E}} = \{F_{X_{1E}}\rho_1 + F_{X_{1\bar{E}}}(1-\rho_1) - F_{X_{0\bar{E}}}(1-\rho_0)\}/\rho_0 \tag{5}$$

Then, the utility function is estimated as

$$\hat{U}(c,r) \propto \hat{F}_{X_{1\bar{E}}}(c)(1-\hat{\rho}_1) - \hat{F}_{X_{0\bar{E}}}(c)(1-\hat{\rho}_0) - r \times \{\hat{F}_{X_{1E}}\hat{\rho}_1 + \hat{F}_{X_{1\bar{E}}}(c)(1-\hat{\rho}_1)\} + A$$

Obviously, the same process could be applied for any of the four marker distributions.

Suppose, for example, that the biomarker values follow a Gaussian distribution in the populations $i\bar{E}$ with mean $\mu_{i\bar{E}}$ and variance $\sigma^2_{i\bar{E}}$, and that $r=0$ (the two treatment options have equal toxicities), then the utility function is expressed as

$$U(c,r) \propto \Phi\left(\frac{c-\mu_{1\bar{E}}}{\sigma_{1\bar{E}}}\right)(1-\rho_1) - \Phi\left(\frac{c-\mu_{0\bar{E}}}{\sigma_{0\bar{E}}}\right)(1-\rho_0) + A$$

where Φ denotes the standard normal cumulative distribution function. In this very special case, an explicit solution exists for the optimal threshold, c^*:

$$c^* = \frac{\mu_{1\bar{E}}(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\sigma^2_{1\bar{E}}\log(b^2R^2)}}{b^2-1}$$

with $a = (\mu_{0\bar{E}} - \mu_{1\bar{E}})$, $b = \frac{\sigma_{0\bar{E}}}{\sigma_{1\bar{E}}}$ and $R = \frac{1-\rho_0}{1-\rho_1}$.

Or, when $\sigma_{0\bar{E}} = \sigma_{1\bar{E}}$

$$c^* = \frac{\sigma_0^2\log(R^2) + \mu_{0\bar{E}} - \mu_{1\bar{E}}}{2a}$$

where σ_0 denotes the standard deviation common to the $0\bar{E}$ and $1\bar{E}$ populations.

The variance of the optimal threshold can be derived using the Delta method. However, it is likely that in most cases there is no explicit solution for the optimal threshold. In such cases, numerical algorithms may be used to

optimize the utility function so as to estimate the threshold.[21,22] In this case, the Delta method cannot be used anymore. One solution is to use the Bayesian inference to derive an estimate and a credible interval of the optimal threshold.

An estimate of the optimal threshold could be, for example, the median of its posterior distribution, and a 95% credible interval could be obtained using the 2.5% and 97.5% quantiles of its posterior distribution. As the optimal threshold is a function of the marker distribution in each subgroup of patients and of the mean risk of event in each treatment group, its posterior distribution can be derived from the posterior distribution of the marker distribution parameters and of the posterior distribution of the mean risks of event using a Monte-Carlo method.[24]

Let $\boldsymbol{\theta}_{ij}$ denote the parameter(s) of the distribution function of the marker or the parameters of the distribution function of the mean risk of event in each treatment group (note that $\boldsymbol{\theta}_{ij}$ may be a vector of parameters). Let also $p(\boldsymbol{\theta}_{ij})$ denote the prior distribution on $\boldsymbol{\theta}_{ij}$. This prior distribution may be non-informative, or vaguely informative,[24] when there is no prior information. Let $\boldsymbol{x}_{ij} = x_{ij1}, \ldots, x_{ijn_{ij}}$ be the set of n_{ij} biomarker measurements in the population ij. The posterior distribution of $\boldsymbol{\theta}_{ij}$ is obtained using the Bayes theorem

$$p(\boldsymbol{\theta}_{ij}|\boldsymbol{x}_{ij}) \propto p(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_{ij})\, p(\boldsymbol{\theta}_{ij})$$

The posterior distribution may not correspond to a known theoretical distribution, except in some special cases. When it is unknown, algorithms are used to sample the parameters from their posterior distribution, such as Markov Chain Monte Carlo (MCMC) algorithms.[24]

When using MCMC algorithms, it is possible to sample K values from the posterior distribution of the parameters and to obtain K sampled functions from the posterior distribution of the utility function. Each of these K functions is then maximized which leads to K values of c^* that are a sample of the posterior distribution of the optimal threshold from which it is possible to extract an estimate of the optimal threshold and a credible interval.

It is worth noting that given the sampled parameters, there may be some of the K iterations for which the optimal threshold is not defined (4); this corresponds to cases where one of the two treatment strategies would always be preferred, no matter the marker value. As the posterior distribution of the optimal threshold is defined only in cases where the optimal threshold exists, it makes sense to exclude these samples a posteriori. Such a situation could occur when the candidate marker has a poor capacity for treatment selection. One limit of this exclusion approach is that, in some cases, it could be time-consuming to construct an MCMC chain of sufficient length.

3 Simulation study

A simulation study was conducted to assess the above estimation method in multiple settings that are defined hereafter.

3.1 Design of the simulations

Let us consider N patients, of whom $N/2$ are randomized to the reference treatment arm and $N/2$ are randomized to the innovative treatment arm. The biomarker values were sampled using Gaussian distributions

$$\boldsymbol{x}_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2_{ij})$$

except for patients who developed the event of interest in the reference treatment arm (population $0E$). For this population, the marker values were sampled from the compound distribution (5), using a Metropolis algorithm,[24] to fulfil the randomization constraint.

Non-informative priors were assumed for the marker distribution parameters: $p(\mu_{0E}, \sigma^2_{0E}) \propto \sigma^2_{0E}$, $p(\mu_{1E}, \sigma^2_{1E}) \propto \sigma^2_{1E}$ and $p(\mu_{1\bar{E}}, \sigma^2_{1\bar{E}}) \propto \sigma^2_{1\bar{E}}$. An MCMC algorithm was used to sample values from the posterior distribution of each parameter. The results of 2000 values from this algorithm were retained leading to 2000 values for each parameter. For each iteration of the MCMC algorithm, the corresponding optimal threshold was calculated using numerical methods (a Newton-type algorithm was used in these simulations using the nlm function of the R software[25]). The median, and 2.5 and 97.5% quantiles of the optimal threshold distribution were calculated.

In the simulation study, multiple settings were considered by varying the following parameters:

- the total number of patients $N = \{500, 1000, 1500, 5000\}$;

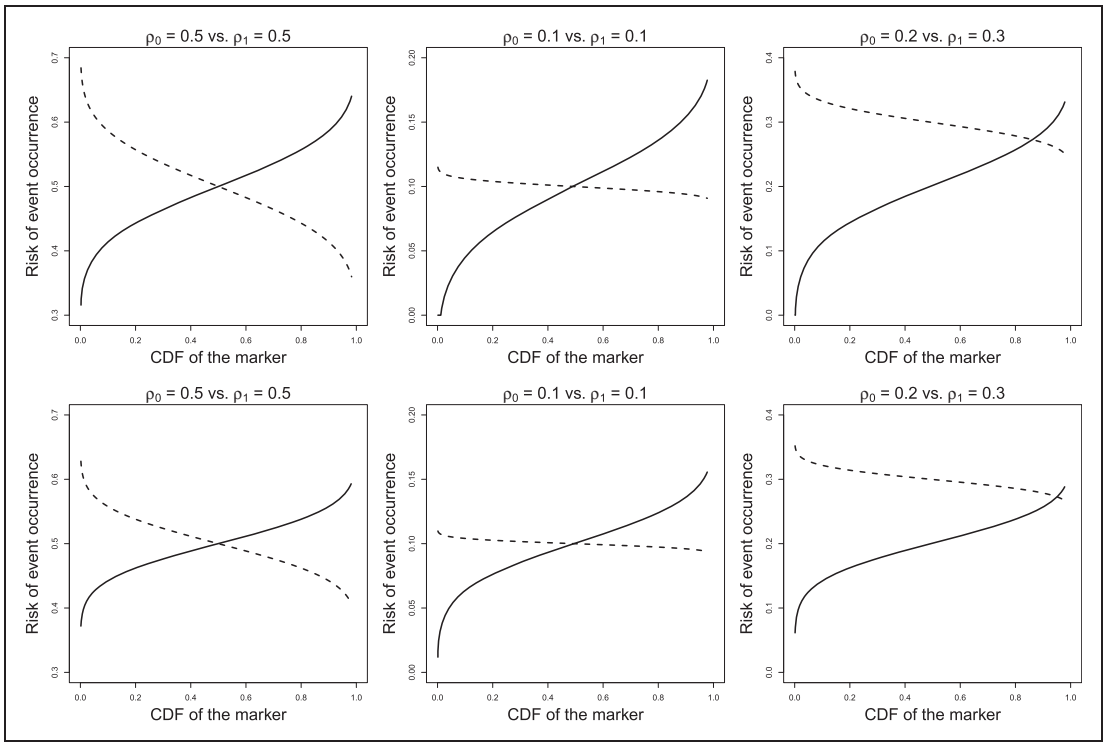


Figure 1. Risk curves according to the cumulative distribution function (CDF) of the marker and the treatments. First row: scenario A (moderate capacity for treatment selection). Second row: scenario B (low capacity for treatment selection). ρ_0: mean risk of event in the reference arm; ρ_1: mean risk of event in the innovative arm.

- the mean risk of event in the two treatment groups ρ_0 vs. $\rho_1 = \{0.5$ vs. $0.5, 0.1$ vs. $0.1, 0.2$ vs. $0.3\}$;
- the r ratio according to the mean risk settings. When ρ_0 vs. $\rho_1 = \{0.5$ vs. $0.5, 0.1$ vs. $0.1\}$, $r = \{0, 0.025, 0.05\}$, and when ρ_0 vs. $\rho_1 = \{0.2$ vs. $0.3\}$, $r = \{0.05, 0.1, 0.15\}$;
- the overall capacity of the marker for treatment selection, by varying the distribution parameters of the marker in each population. We call scenario A the scenario with a moderate capacity and scenario B the scenario with a low capacity. Figure 1 presents the risk curves associated with each setting in order to visualize the capacity of the marker for treatment selection in each scenario.

For each setting, 2000 datasets were simulated and four criteria were used to evaluate the estimation method:

- the normalized mean bias (NMB) of the estimate defined as the difference between the expected value of the estimated optimal threshold and its theoretical value, divided by the variance of the estimate;
- the coverage probability of the 95% credible interval defined by the proportion of simulations for which the credible interval contained the theoretical value;
- the mean width of the 95% credible interval which assesses the precision of the estimation method;
- the proportion of simulations for which at least 30% iterations of the MCMC algorithm failed to estimate the optimal threshold, denoted by γ_{30}.

The latter criterion is an indicator of the ability of the proposed method to detect and estimate the optimal threshold. A high proportion would be a strong indication that the marker would have a poor capacity for treatment selection, and consequently that the search for an optimal threshold may be irrelevant.

3.2 Results of the simulation study

Tables 1 and 2 present the results of the simulation study for scenarios A and B, respectively.

In almost all settings and scenarios, an increase of N was associated with a decrease in the NMB, a decrease of the mean width of the credible interval and a coverage probability of the credible interval closer to 95%.

Table 1. Normalized mean bias (NMB) of the optimal threshold estimate, together with coverage probability (CP) and mean width of the credible interval (WCI) of this estimate for various parameter settings under scenario A.

ρ_0 vs. ρ_1	r ratio	N	NMB	CP (%)	WCI	γ_{30}
0.5 vs. 0.5	0	500	−0.024	94.9	4.17	0.07
		1000	−0.006	94.7	2.88	0.01
		1500	0.012	95.2	2.27	0.00
		5000	−0.103	93.7	1.17	0.00
	0.025	500	0.026	94.9	4.19	0.08
		1000	0.024	94.6	2.92	0.03
		1500	0.035	94.9	2.31	0.01
		5000	0.031	94.0	1.20	0.00
	0.05	500	0.082	94.8	4.19	0.12
		1000	0.061	94.1	3.00	0.06
		1500	0.064	94.8	2.41	0.02
		5000	0.005	94.3	1.25	0.00
0.1 vs. 0.1	0	500	0.000	91.9	6.81	0.29
		1000	−0.019	93.1	6.13	0.22
		1500	−0.030	92.9	5.68	0.17
		5000	−0.053	94.7	3.93	0.05
	0.025	500	0.231	91.1	6.82	0.31
		1000	0.176	91.7	5.92	0.27
		1500	0.128	92.0	5.42	0.24
		5000	0.048	93.6	3.89	0.17
	0.05	500	0.493	87.1	6.37	0.38
		1000	0.447	88.9	5.55	0.38
		1500	0.411	90.0	5.12	0.36
		5000	0.210	91.0	3.98	0.33
0.2 vs. 0.3	0.05	500	−0.135	93.2	6.06	0.25
		1000	−0.130	93.4	4.64	0.20
		1500	−0.137	93.7	3.96	0.17
		5000	−0.084	94.7	2.27	0.05
	0.1	500	0.026	92.0	6.05	0.18
		1000	0.026	93.8	4.62	0.10
		1500	0.003	94.3	3.86	0.05
		5000	−0.093	93.8	1.99	0.00
	0.15	500	0.195	91.7	5.60	0.25
		1000	0.182	92.6	4.39	0.20
		1500	0.138	93.6	3.80	0.17
		5000	−0.038	93.6	2.24	0.05

In scenario A, the NMB was always close or lower than 10% when $\rho_0 = 0.5$ and $\rho_1 = 0.5$, whatever the r ratio. In other mean risk settings, the NMB was always lower than 10% when the r ratio was equal to the difference in the mean risks of event between the two treatment arms. Moreover, it was lower than 10% with other r ratios when N was high (at least 5000).

When $\rho_0 = 0.5$ and $\rho_1 = 0.5$, the coverage probability of the estimated credible interval was within (93.7%; 95.2%) in all settings. When reducing the proportion of events in each arm with $\rho_0 = 0.1$ and $\rho_1 = 0.1$, at least 1000 patients were needed with $r = 0$ (which corresponds to the difference between ρ_0 and ρ_1) for the coverage probability to be close or greater than 93%. When $r \neq 0$, at least 5000 patients were needed for the coverage probability to be greater than 93%. With intermediate mean risks of event $\rho_0 = 0.2$ and $\rho_1 = 0.3$, at least 1000 patients were needed with $r = 0.1$ (which corresponds to the difference between ρ_0 and ρ_1) for the coverage probability to be greater than 93%. When $r = 0.15$, at least 1500 patients were needed for the coverage probability to be greater than 93%. When $r = 0.05$, at least 500 patients were needed for the coverage probability to be greater than 93%.

In scenario B, with a treatment selection marker that has a poorer capacity to guide treatment decisions than in scenario A, the same trends were observed except that the NMB in the threshold estimate was higher, and that

Table 2. Normalized mean bias (NMB) of the optimal threshold estimate, together with coverage probability (CP) and mean width of the credible interval (WCI) of this estimate for various parameter settings under scenario B.

ρ_0 vs. ρ_1	*r* ratio	N	NMB	CP (%)	WCI	γ_{30}
0.5 vs. 0.5	0	500	−0.019	93.6	5.66	0.20
		1000	−0.002	94.1	4.32	0.09
		1500	0.009	94.4	3.55	0.03
		5000	0.026	94.3	1.81	0.00
	0.025	500	0.075	94.1	5.67	0.22
		1000	0.074	93.4	4.34	0.12
		1500	0.081	94.6	3.60	0.07
		5000	0.030	94.2	1.90	0.00
	0.05	500	0.177	93.6	5.57	0.26
		1000	0.175	92.9	4.33	0.17
		1500	0.173	94.3	3.68	0.13
		5000	0.041	94.1	2.09	0.02
0.1 vs. 0.1	0	500	−0.005	91.9	7.46	0.34
		1000	−0.015	91.9	7.01	0.32
		1500	−0.007	92.5	6.55	0.25
		5000	−0.010	94.2	5.20	0.14
	0.025	500	0.166	92.3	7.48	0.35
		1000	0.084	91.7	6.67	0.37
		1500	0.050	92.2	6.16	0.32
		5000	−0.194	92.5	4.87	0.32
	0.05	500	0.801	80.1	6.87	0.43
		1000	0.770	82.8	6.12	0.47
		1500	0.807	83.5	5.66	0.45
		5000	0.808	88.7	4.80	0.49
0.2 vs. 0.3	0.05	500	−0.256	90.7	7.04	0.36
		1000	−0.285	91.1	5.83	0.33
		1500	−0.266	92.7	5.22	0.32
		5000	−0.243	94.0	3.51	0.23
	0.1	500	0.031	91.9	7.19	0.29
		1000	0.001	92.5	5.98	0.21
		1500	−0.001	92.2	5.34	0.16
		5000	−0.067	93.4	3.14	0.03
	0.15	500	0.336	90.2	6.59	0.36
		1000	0.314	91.0	5.46	0.33
		1500	0.266	92.1	4.93	0.31
		5000	0.183	92.9	3.40	0.22

more patients were needed for the coverage probability to be close to 95% in some settings. Overall, a high γ_{30} was associated with higher NMB in almost all settings.

4 Example: The PETACC-8 trial

4.1 Context

The PETACC-8 trial is an open-label, randomized, controlled, multinational phase III study in patients aged 18 to 75 years with pathologically confirmed and resected stage III colon adenocarcinoma.[26] The trial compares the efficacy of FOLFOX4 (oxaliplatin, fluorouracil and leucovorin) to that of FOLFOX4 + cetuximab (an EGFR inhibitor). As patients with KRAS exon 2 mutated tumours were found resistant to EGFR antibodies in the metastatic setting,[3] an amendment to the protocol restricted the enrolment to patients with KRAS wild-type tumours. For the identification of a treatment selection marker, the analysed outcome was cancer recurrence or death within 21 months.

Herein, the aim was to estimate the optimal threshold of the DDR2 gene expression level as it was found to be a potential treatment selection marker in this study. Accordingly, analysis was restricted to patients who had a

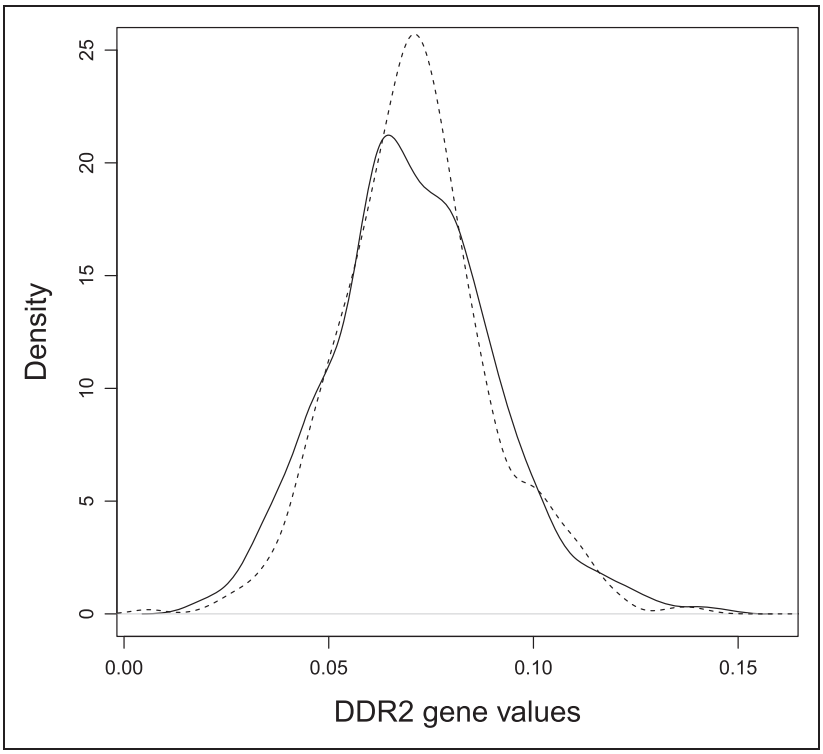


Figure 2. Comparison of the DDR2 distributions in both treatment arms. Solid line: FOLFOX-4 arm; dotted line: FOLFOX-4 + cetuximab arm.

measure of DDR2 gene expression, and thus 1068 patients were analysed. As such a selection could lead to imbalance in marker distribution between treatment groups, the empirical distribution of the DDR2 gene expression in both treatment arms was investigated and found to be quite similar (Figure 2).

The 21-month post-treatment follow-up period was a good compromise between the number of censored data and a sufficient interval of time to observe treatment effects (without censoring, the follow-up would have been too short). The number of censored data was 60, and outcomes for these 60 patients were imputed. In the imputation method used,[27] the probabilities of event occurrence were estimated using the uncensored data at 21 months of follow-up with a logistic model that included the main prognostic factors of the study: the treatment arm, information about the bowel (obstructed/perforated or not), the histopathological grade, the pT pathological stage, and the pN pathological stage. After estimating the model parameters, the probabilities of event occurrence in patients with censored data were predicted and their statuses sampled from a Bernoulli law.

At 21 months post-treatment, the proportion of events (including the results of outcome imputation) in each arm was estimated to be $\hat{\rho}_0 = 0.16$ in the FOLFOX-4 arm (denoted as the reference arm), and $\hat{\rho}_1 = 0.17$ in the FOLFOX-4 + cetuximab arm (denoted as the innovative arm).

4.2 Application

The marker distributions in populations $0\bar{E}$, $1E$ and $1\bar{E}$ could easily be approximated by theoretical distributions; however, it was harder to fit a theoretical distribution in population $0E$. As the randomization constraint allows us to fit three of the four marker distributions, we decided to estimate indirectly the marker distribution in patients who received the reference treatment and presented the event. Theoretical distributions were fitted to the three remaining marker distributions:

- Population $0\bar{E}$: $\boldsymbol{x}_{0\bar{E}} \sim \mathcal{N}(\mu_{0\bar{E}}, \sigma_{0\bar{E}})$ with $\mu_{0\bar{E}} = 0.0714$ and $\sigma_{0\bar{E}} = 0.0200$;
- Population $1E$: $\boldsymbol{x}_{1E} \sim \mathcal{T}(\mu_{1E}, \sigma_{1E}, \mathrm{df}_{1E})$ with $\mu_{1E} = 0.0733$, $\sigma_{1E} = 0.0169$ and $\mathrm{df}_{1E} = 6.294$;
- Population $1\bar{E}$: $\log(\boldsymbol{x}_{1\bar{E}}) \sim \mathcal{N}(\mu_{1\bar{E}}, \sigma_{1\bar{E}})$ with $\mu_{1\bar{E}} = -2.6806$ and $\sigma_{1\bar{E}} = 0.2614$;

with $\mathcal{T}$ denoting a Student-t distribution.

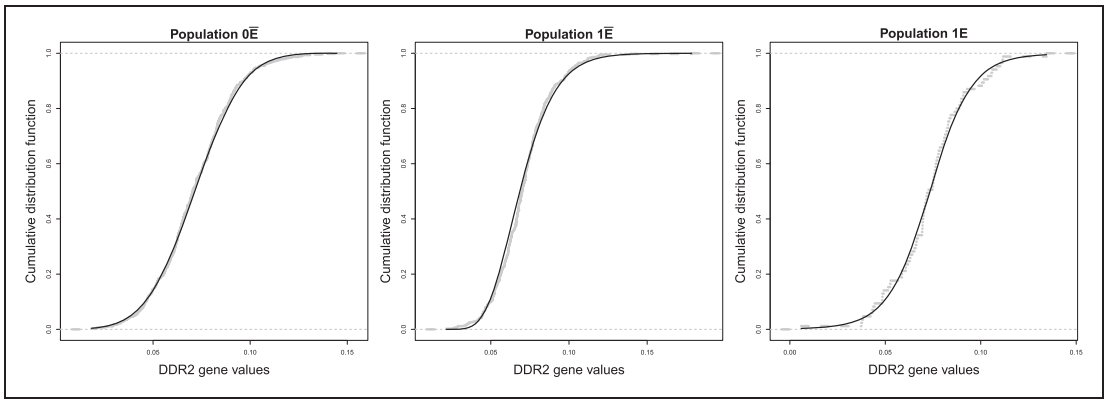


Figure 3. Theoretical cumulative distributions fitted on empirical cumulative distributions. Black solid lines: theoretical distributions; grey dotted lines: empirical distributions.

Figure 3 presents the three theoretical cumulative distributions fitted on the empirical cumulative distributions of the marker for each group. The empirical distributions are well fitted by the defined theoretical distributions. An MCMC algorithm was used to sample values from the posterior distribution of the parameters for the population $1E$ as it did not correspond to a known theoretical distribution. Three MCMC chains were built for each distribution parameter, each one with a length of 5000 values and a thin equal to three to reduce autocorrelation. The potential scale reduction factors[24] were checked for each parameter of the Student distribution and were all equal to 1.

The distribution parameters in population $0\bar{E}$ and population $1\bar{E}$ were sampled, from their explicit posterior distribution, using non-informative priors. The mean risks of event in both arms were sampled in a Beta distribution using Jeffrey's prior.[24]

The utility function was maximized for each set of sampled parameters using three different r ratios: 0, 0.01 and 0.02. The estimates of the optimal threshold, their 95% credible interval and the proportion of events in each arm under the marker-based treatment selection are presented in Table 3 and defined as:

- Proportion of events under the reference treatment arm when the marker selects this treatment option: $P(E = 1|T = 0, X > c^*)$
- Proportion of events under the innovative arm when the marker selects this treatment option: $P(E = 1|T = 1, X \leq c^*)$.

The optimal threshold estimate is almost the same for the three r ratios and their credible intervals superimpose. The method gives a precise estimate of the optimal threshold with narrow credible intervals. These credible intervals are precise as the marker values range between 0.006 and 0.178. The proportion of events under the marker-based treatment selection in each treatment arm was lower as compared to the mean proportion of events that were initially obtained in the trial. The proportion of sampled parameters in each MCMC chain for which there was no optimal threshold was equal to zero in all considered settings.

Figure 4 presents the relative utility of using the DDR2 gene for treatment selection. This figure shows that the range of r that was considered is relevant as it corresponds to a utility of the marker-based strategy greater than 0.

A patient with a marker value above the optimal threshold estimate is likely to benefit greater from the FOLFOX4 therapy alone, while a patient with a marker value below the optimal threshold is likely to benefit from the combination of FOLFOX4 + cetuximab.

While there was not a significant difference in mean risks of event of each treatment arm, the method developed herein shows that using the DDR2 gene to guide treatment selection could improve the overall health condition of the target population by decreasing the risk of event in both treatment arms. However, Figure 4 shows that better markers may be used as the DDR2 gene expression level is still far from the perfect treatment selection marker.

5 Discussion

In the present study, the Bayesian inference was used to estimate the optimal threshold of a quantitative treatment selection marker and its credible interval. The proposed method relies on modelling the marker distribution in

Table 3. Optimal threshold estimates and credible intervals for different *r* ratios.

r ratio	Optimal threshold	$P(E = 1 \mid T = 0, X > c^*)$	$P(E = 1 \mid T = 1, X \leq c^*)$	Relative utility
0	0.073 (0.068; 0.078)	0.123	0.145	0.308
0.01	0.072 (0.067; 0.078)	0.127	0.148	0.275
0.02	0.071 (0.067; 0.077)	0.125	0.147	0.242

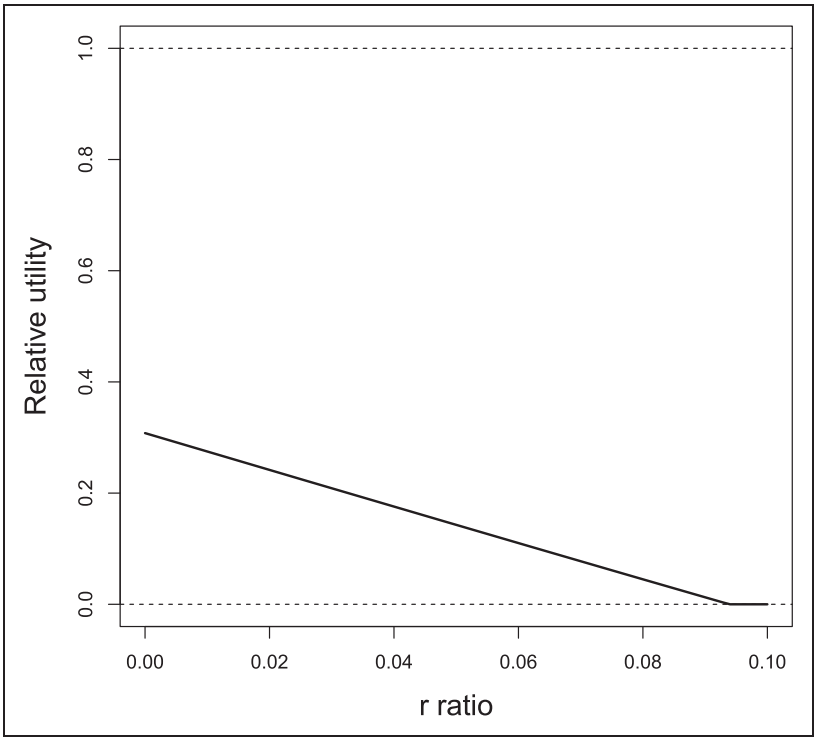


Figure 4. Relative utility curve of the DDR2 gene. On the Y-axis, 0 corresponds to a useless treatment selection marker, while 1 corresponds to a perfect treatment selection marker.

subgroups of patients (those who developed the event of interest or not in both treatment arms) and expressing a utility function that quantifies the health condition in the target population when using the marker to guide treatment selection. This method can handle any type of distribution which offers a flexible tool to estimate the optimal threshold of a treatment selection marker.

The simulation study showed that, in some settings (mean risk of event in each arm equal to 0.5, $r = 0$), an $N = 500$ is sufficient for the method to have low bias (less than 10%) and coverage probability close to 95%; however, in other settings, it may need N higher than 5000 patients to have good properties. A large N is needed when the mean risks of event are low (as it means that only few patients experience the event of interest) and when r is far from the difference in mean risks of event. The need for large samples is not directly due to the proposed method, but rather to treatment selection markers; it is well known that a large number of patients is needed to detect a significant interaction between treatments and a marker,[28] and even more patients are needed to estimate precisely a threshold. Furthermore, in the present study, Gaussian distributions were defined in the simulation settings for the marker in the three groups to illustrate the properties of the method and to highlight the need for high sample sizes, but investigations are needed for other distributions.

To the best of our knowledge, estimating the optimal threshold on the marker scale has yet to be reported, while this is the scale clinicians need in practice. Many authors chose to express the optimal threshold on the difference in risks scale[5–7,11,18] and to estimate it using risk prediction models. Although it is possible to derive an expression of the optimal threshold estimate on the marker scale with these methods, it is harder to obtain a confidence interval. The Bayesian method presented herein provides an easy way to estimate the credible interval of the optimal threshold estimate. We demonstrated the relationship between the expression of the utility function proposed

herein and those reported by other authors; the main difference is that it does not rely on risk modelling, but on the marker distribution modelling. We also consider that it is easier to model the marker distribution in subgroups of patients than modelling the risk of event given the treatments and the marker values.

The main difficulty of the method relies on modelling correctly the marker distribution in each subgroup of patients. Multiple tools exist to help users choose adequate theoretical distributions to fit the data.[14–16] The use of Dirichlet processes[24,29,30] to model the cumulative distribution function of the marker in each subgroup could also be considered and could help users when no theoretical distribution fits the data correctly.

The proposed method relies on the Bayesian inference. In the example of application presented herein, non-informative priors were used, but informative ones could also be used whenever information on the parameters is available. This is not the first time that this kind of method is used to estimate the solution of an optimization problem,[21–23] especially in the context of diagnostic or prognostic marker analysis. However, it is the first time this has been applied to the estimation of the optimal threshold of treatment selection markers, which is a more complex task than in the diagnostic and prognostic setting. In the simulation study herein, we proposed to use the γ_{30} indicator to detect situations when there is no optimal threshold to estimate. Instead of modelling separately the three marker distributions, it could be possible to model them jointly including the constraint that the optimal threshold must exist, but these kinds of constraints are not easy to include. The proposed solution was to exclude the simulations for which no optimal threshold could be estimated, but in some situations this can be a time-consuming approach in order to obtain a sufficient number of samples from the optimal threshold posterior distribution. The finding that a high number of iterations in the MCMC process failed to estimate an optimal threshold (reflected by the γ_{30} indicator) is a strong indication that the marker may have a poor capacity for treatment selection, which means that the marker-based strategy is not better than one of the two extreme strategies. However, it is of note that, as the utility function is maximized numerically, the choice of the optimization algorithm is likely to impact the optimal threshold estimation. In rare situations, the optimal threshold could not be defined because of optimization algorithm failures. Here, the Newton-type algorithm of the nlm function of R[25] was used to maximize the utility function; however, algorithms with better performance could be considered to improve the results in some settings (some algorithms allow box constraints on the marker values that could be used to be more effective, many other algorithms are available in the optim function of R).

Several assumptions are made in Section 2 to simplify the expression of the utility function by summing up the four utilities in only one quantity: the r ratio. These assumptions are made for didactic purpose, which means that it is still possible to estimate an optimal threshold without making these assumptions. For example, if information is available to estimate each utility quantity separately then there is no need to make the Vickers et al.[5] assumptions. More generally, the utility function given in equation (1) can be applied in all situations where the outcome is a binary event (success or failure).

An extension of decision curves and its connection with the one proposed by Huang et al.[18] are presented. This kind of curve may be useful to define a relevant range of r ratios and avoid using extreme values for which the optimal threshold may not be defined. Moreover, these give information about the performance that new markers could achieve by comparing their utility to that of the perfect marker.

One limit of the proposed method is that the outcome of all patients must be known which is rarely the case in most studies; herein, the censored outcomes were imputed. The proposed method may be extended using methods already developed for correcting the bias induced in sensitivity and specificity estimates with censored data[31] in order to avoid using imputation methods.

Janes et al.[28] already highlighted the need of large samples to evaluate the benefit of a biomarker for selecting patient treatment. The same conclusions were found in the present study, as the results highlight the need of large samples to estimate the optimal threshold of treatment selection markers. That is why more developments would be needed to provide simple ways of estimating the number of subjects needed to estimate with a sufficient precision the optimal threshold.

Acknowledgements

The authors thank Dr Philip Robinson (Hospices Civils de Lyon) whose suggestions improved significantly the present manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Fondation pour la Recherche Médicale (FRM grant number ECO20160736050 to YB).

ORCID iD

Yoann Blangero http://orcid.org/0000-0003-0309-3399

Supplemental material

Supplemental material is available for this article online.

References

1. Ballman K. Biomarker: predictive or prognostic?. *J Clin Oncol* 2015; **33**: 3968–3971.
2. Italiano A. Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol* 2011; **29**: 4718.
3. Lièvre A, Bachet J, Boige V, et al. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *J Clin Oncol* 2008; **26**: 374–379.
4. Byar D. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Stat Med* 1985; **4**: 255–263.
5. Vickers A, Kattan M and Sargent D. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* 2007; **8**: 14.
6. Janes H, Pepe M, Bossuyt P, et al. Measuring the performance of markers for guiding treatment decisions. *Ann Intern Med* 2011; **154**: 253–259.
7. Janes H, Brown M, Huang Y, et al. An approach to evaluating and comparing biomarkers for patient treatment selection. *Int J Biostat* 2014; **10**: 99–121.
8. Huang Y, Gilbert P and Janes H. Assessing treatment-selection markers using a potential outcomes framework. *Biometrics* 2012; **68**: 687–696.
9. Zhang Z, Nie L, Soon G, et al. The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. *Ann Appl Stat* 2014; **8**: 2336–2355.
10. Sox H, Higgins M and Owens D. *Medical decision making*. 2nd ed. Chichester: John Wiley & Sons, 2013.
11. Janes H, Pepe M and Huang Y. A framework for evaluating markers used to select patient treatment. *Med Decis Making* 2014; **34**: 159–167.
12. Collins G and Moons K. Comparing risk prediction models. *BMJ* 2012; **344**: e3186.
13. Van Calster B and Vickers A. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making* 2015; **35**: 162–169.
14. Cullen A and Frey H. *Probabilistic techniques in exposure assessment*. New York: Plenum Press, 1999.
15. Forbes C, Evans M, Hastings N, et al. *Statistical distributions*. 4th ed. Hoboken: John Wiley & Sons, 2000.
16. Delignette-Muller ML and Dutang C. fitdistrplus: an R package for fitting distributions. *J Stat Softw* 2015; **64**: 1–34.
17. Vickers AJ and Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**: 565–74.
18. Huang Y, Laber E and Janes H. Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics* 2015; **16**: 383–399.
19. Baker SG, Cook NR, Vickers A, et al. Using relative utility curves to evaluate risk prediction. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 729–748.
20. Efron B and Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1986; **1**: 54–75.
21. Schisterman E and Perkins N. Confidence intervals of the Youden index and corresponding optimal cut-point. *Commun Stat Simul Comput* 2007; **36**: 549–563.
22. Subtil F and Rabilloud M. A Bayesian method to estimate the optimal threshold of a longitudinal biomarker. *Biom J* 2010; **52**: 333–347.
23. Subtil F and Rabilloud M. Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC Med Inform Decis Making* 2014; **14**: 53.
24. Gelman A, Carlin J, Stern H, et al. *Bayesian data analysis*. 3rd ed. Boca Raton: CRC Press, 2014.
25. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017.
26. Taieb J, Tabernero J, Mini E, et al. Oxaliplatin, fluorouracil, and leucovorin with or without cetuximab in patients with resected stage III colon cancer (PETACC-8): an open-label, randomised phase 3 trial. *Lancet Oncol* 2014; **15**: 862–873.
27. Dmitrienko A, Molenberghs G, Chuang-Stein C, et al. *Analysis of clinical trials using SAS: a practical guide*. Cary: SAS Institute, 2005.

28. Janes H, Brown M and Pepe M. Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Stat Med* 2015; **34**: 3503–3515.
29. Escobar M. Estimating normal means with a Dirichlet process prior. *J Am Stat Assoc* 1994; **89**: 268–277.
30. Ohlssen D, Sharples L and Spiegelhalter D. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat Med* 2007; **26**: 2088–2112.
31. Blanche P, Dartigues JF and Jacqmin-Gadda H. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biom J* 2013; **55**: 687–704.

3.1.3 Principaux résultats de l'article

Les résultats des simulations montrent que dans certains scénarios ($\rho_0 = \rho_1 = 0.5$ et $\frac{C_Z}{C_Y} = 0$), un effectif total de 500 patients est suffisant pour que la méthode Bayésienne d'estimation du seuil optimal ait un biais faible et une probabilité de couverture proche de 95 % ; cependant, dans d'autres configurations, il peut être nécessaire d'avoir un effectif total supérieur à 5000 patients pour avoir de bonnes propriétés. Une taille d'échantillon très grande est nécessaire quand le risque moyen de survenue d'évènement est faible (car cela signifie que peu de patients vont développer l'évènement d'intérêt et que la distribution du marqueur pour ces patients est mal estimée) et quand $\frac{C_Z}{C_Y}$ est éloigné de la différence de risques moyenne ($\rho_1 - \rho_0$). Le besoin de grands échantillons n'est pas directement dû à la méthode proposée, mais plutôt aux problématiques d'analyse de marqueurs prédictifs d'une part et de recherche de seuil d'autre part ; il est bien connu qu'un grand nombre de patients est nécessaire pour détecter ce type de marqueur (Janes et al., 2015a). Par ailleurs, l'estimation d'un seuil optimal, même dans le cadre d'un simple test diagnostique, nécessite également un nombre élevé de patients (Jund et al., 2005). La combinaison de ces deux contextes aboutit à des effectifs nécessaires très élevés.

La méthode a été appliquée au niveau d'amplification du gène *DDR2* mesuré dans le cadre de l'essai PETACC-8. L'identification du seuil optimal de ce marqueur a permis de modifier la règle d'allocation des traitements et de passer d'un risque moyen de survenue d'évènement dans le bras FOLFOX4 de 16 % à 12.3 %, et d'un risque moyen dans le bras FOLFOX4 + cetuximab de 17 % à 14.5 %, illustrant ainsi l'intérêt de la méthode en pratique clinique. Dans le cadre de cet essai, les profils de toxicité des traitements étaient relativement proches, ce qui justifiait l'utilisation d'un ratio $\frac{C_Z}{C_Y}$ faible pour l'estimation du seuil optimal du marqueur.

3.2 Compléments à l'article

3.2.1 Solution explicite pour le seuil optimal

Dans l'article publié, les simulations évaluant les performances de la méthode d'estimation ont nécessité la détermination du seuil optimal théorique pour les différents scénarios envisagés.

En reprenant l'exemple d'un marqueur dont les valeurs suivent une distribution Gaussienne dans les groupes de patients $z\bar{Y}$ (les patients n'ayant pas développé l'évènement dans chacun des deux bras de traitement) avec pour moyennes $\mu_{z\bar{Y}}$ et variances $\sigma^2_{z\bar{Y}}$, et en supposant également que $\frac{C_Z}{C_Y} = 0$ (les traitements ont des toxicités équivalentes), alors la fonction d'utilité s'écrit sous la forme suivante :

$$U(c) \propto \Phi\left(\frac{c - \mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right) \times (1 - \rho_1) - \Phi\left(\frac{c - \mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right) \times (1 - \rho_0),$$

avec $\Phi(.)$ la fonction de répartition d'une loi normale centrée réduite. Dans ce cas, il existe une

solution explicite, c^*, pour le seuil optimal (la démonstration est donnée en Annexe D) :

$$c^* = \frac{\mu_{0\bar{Y}}(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\sigma_{0\bar{Y}}^2 \log(b^2R^2)}}{b^2-1},$$

avec $a = (\mu_{1\bar{Y}} - \mu_{0\bar{Y}})$, $b = \frac{\sigma_{1\bar{Y}}}{\sigma_{0\bar{Y}}}$, et $R = \frac{1-\rho_0}{1-\rho_1}$.

Lorsque $\sigma_{0\bar{Y}}^2 = \sigma_{1\bar{Y}}^2 = \sigma_{\bar{Y}}^2$ alors l'expression du seuil optimal est simplifiée :

$$c^* = \frac{\sigma_{\bar{Y}}^2 \log(R^2) + \mu_{1\bar{Y}}^2 - \mu_{0\bar{Y}}^2}{2a}.$$

L'estimation du seuil optimal repose sur les estimateurs de a, b, R, $\mu_{0\bar{Y}}$, $\mu_{1\bar{Y}}$, $\sigma_{0\bar{Y}}$, $\sigma_{1\bar{Y}}$, ρ_0 et ρ_1 :

$$\widehat{c}^* = \frac{\widehat{\mu}_{0\bar{Y}}(\widehat{b}^2-1) - \widehat{a} + \widehat{b}\sqrt{\widehat{a}^2 + (\widehat{b}^2-1)\widehat{\sigma}_{0\bar{Y}}^2 \log(\widehat{b}^2\widehat{R}^2)}}{\widehat{b}^2-1}.$$

Ainsi, dans l'algorithme MCMC défini pour estimer le seuil optimal, la maximisation de la fonction d'utilité par une méthode numérique à chaque itération de l'algorithme peut être remplacée par l'utilisation de cette solution explicite. Dans le cas très précis évoqué précédemment, l'existence d'une solution explicite pour le seuil optimal permet d'envisager une alternative à l'algorithme MCMC. Celle-ci repose sur le théorème central limite permettant d'approximer la distribution de l'estimateur du seuil optimal par une loi normale, et dont la variance peut être calculée de manière analytique à l'aide de la méthode Delta (les détails du calcul sont donnés en Annexe D). En posant :

$$W = a^2 + \left(b^2-1\right)\sigma_{0\bar{Y}}^2 \log\left(\frac{\sigma_{1\bar{Y}}^2(1-\rho_0)^2}{\sigma_{0\bar{Y}}^2(1-\rho_1)^2}\right),$$

alors il est possible d'écrire explicitement $\mathrm{Var}(\widehat{c}^*)$ de la manière suivante :

$$\begin{aligned}
\mathrm{Var}(\widehat{c}^*) \approx & \left\{\frac{-1 + ba \times W^{-1/2}}{b^2-1}\right\}^2 \times \frac{\sigma_{1\bar{Y}}^2}{n_{1\bar{Y}}} + \left\{\frac{b^2 - ba \times W^{-1/2}}{b^2-1}\right\}^2 \times \frac{\sigma_{0\bar{Y}}^2}{n_{0\bar{Y}}} \\
& + \left\{\frac{2ab + (-b^2-1)\sqrt{W}}{(b^2-1)^2\sigma_{0\bar{Y}}} + \frac{\sigma_{1\bar{Y}} b \times \left[\log(R^2) + 1 - b^{-2}\right]}{(b^2-1)\sqrt{W}}\right\}^2 \times \frac{\sigma_{1\bar{Y}}^2}{2(n_{1\bar{Y}}-1)} \\
& + \left\{\frac{-2ab^2 + b(b^2+1)\sqrt{W}}{(b^2-1)^2\sigma_{0\bar{Y}}} - \frac{\sigma_{0\bar{Y}} b \times \left[\log(R^2) + b^2 - 1\right]}{(b^2-1)\sqrt{W}}\right\}^2 \times \frac{\sigma_{0\bar{Y}}^2}{2(n_{0\bar{Y}}-1)} \\
& + \left\{\frac{b(\sigma_{0\bar{Y}}^2 - \sigma_{1\bar{Y}}^2)}{(b^2-1)(1-\rho_0)\sqrt{W}}\right\}^2 \times \frac{\rho_0(1-\rho_0)}{n_0} + \left\{\frac{b(\sigma_{1\bar{Y}}^2 - \sigma_{0\bar{Y}}^2)}{(b^2-1)(1-\rho_1)\sqrt{W}}\right\}^2 \times \frac{\rho_1(1-\rho_1)}{n_1}.
\end{aligned}$$

Cette variance peut être estimée simplement en utilisant les estimateurs de W, a, b, R, $\mu_{0\bar{Y}}$, $\mu_{1\bar{Y}}$, $\sigma_{0\bar{Y}}$, $\sigma_{1\bar{Y}}$, ρ_0 et ρ_1.

Les propriétés de la méthode bayésienne d'estimation du seuil n'ont pas été comparées à

celles de cette méthode, car cette dernière ne peut s'appliquer que dans des cas très limités. Lorsqu'il n'existe pas de solution explicite du seuil optimal, alors le seuil optimal théorique pour les simulations a été déterminé par des méthodes numériques de maximisation de fonction.

3.2.2 Lien entre la fonction d'utilité et l'expression du bénéfice moyen

Dans plusieurs articles, certains auteurs font référence à la mesure du bénéfice moyen du marqueur pour une règle de décision donnée (Janes et al., 2014b,a; Huang et al., 2015). Le bénéfice moyen peut se définir comme la différence d'utilité entre la stratégie ne reposant pas sur l'évaluation du marqueur pour choisir le traitement, et celle qui consiste à utiliser le marqueur. On considère le cas où la stratégie par défaut (sans utiliser les informations du marqueur) consiste à traiter tous les patients avec le traitement innovant (car $\rho_1 < \rho_0$). Définissons $\delta_W(X)$ la différence de risque entre les deux bras de traitements conditionnellement aux valeurs de X, différence dans laquelle chacun des risques est pondéré par le coût de l'événement pour le traitement en question :

$$\delta_w(X) = \rho_1(X)[C_Y(X) + C_{++}(X)] - \rho_0(X)C_Y(X)\,,$$

alors l'expression générale du bénéfice moyen est définie comme la différence entre les coûts moyens de la stratégie par défaut et de la stratégie de traitement basée sur le marqueur. Le coût moyen de la stratégie par défaut peut s'exprimer sous la forme :

$$\begin{aligned} B_1 = \int_{-\infty}^{+\infty} & \mathrm{P}(Y=1|Z=1,X=x)[C_Y(x) + C_{++}(x) + C_Z(x)] \\ & + \mathrm{P}(Y=0|Z=1,X=x)C_Z(x)\; dF(x)\,. \end{aligned}$$

Concernant la stratégie basée sur le marqueur, son coût moyen peut s'exprimer comme :

$$\begin{aligned} B_X = & \int_{-\infty}^{+\infty} \mathrm{P}(Y=1|Z=0,X=x)C_Y(x)[1-g(x)]\; dF(x) \\ & + \int_{-\infty}^{+\infty} \mathrm{P}(Y=0|Z=1,X=x)C_Z(x)g(x)\; dF(x) \\ & + \int_{-\infty}^{+\infty} \mathrm{P}(Y=1|Z=1,X=x)[C_Y(x) + C_{++}(x) + C_Z(x)]g(x)\; dF(x) + C_M\,, \end{aligned}$$

où $g(X)$ est la règle d'allocation de traitement basée sur le marqueur prenant comme valeur 1 lorsque le traitement innovant est recommandé et 0 dans le cas contraire, et C_M est le coût lié à la mesure du marqueur étudié (Janes et al., 2014b). Le bénéfice moyen de Janes et al. (2014b) s'exprime alors sous la forme suivante :

$$\begin{aligned} \Theta &= B_1 - B_X \\ &= \mathbf{E}[\delta_w(X) + C_Z(X)|g(X)=0] \times \mathrm{P}(g(X)=0) - C_M\,, \end{aligned} \tag{3.2}$$

En faisant l'hypothèse que des valeurs hautes de marqueur sont associées à un meilleur bénéfice du traitement de référence (hypothèse de monotonicité de la différence de risques en fonction de X), alors les auteurs proposent une expression de $g(X)$ qui maximise l'expression du bénéfice (3.2), notée $g^*(X)$, et donnée par :

$$g^*(X) = \mathbb{1}[\delta_w(X) \leq C_Z(X)]\,. \tag{3.3}$$

Il est possible de démontrer que la fonction d'utilité présentée dans l'équation (3.1) est une transformation linéaire du bénéfice moyen (3.2) :

$$\begin{aligned}
U(c) &= \int_c^{+\infty} \mathrm{P}(Y=1,X=x|Z=0)\times U_{0Y}(x) + \mathrm{P}(Y=0,X=x|Z=0)\times U_{0\bar{Y}}(x)\; dx \\
&\quad + \int_{-\infty}^{c} \mathrm{P}(Y=1,X=x|Z=1)\times U_{1Y}(x) + \mathrm{P}(Y=0,X=x|Z=1)\times U_{1\bar{Y}}(x)\; dx \\
&= \int_c^{+\infty} f(x)[\mathrm{P}(Y=1|X=x,Z=0)\times U_{0Y}(x) + \mathrm{P}(Y=0|X=x,Z=0)\times U_{0\bar{Y}}(x)]\; dx \\
&\quad + \int_{-\infty}^{c} f(x)[\mathrm{P}(Y=1|X=x,Z=1)\times U_{1Y}(x) + \mathrm{P}(Y=0|X=x,Z=1)\times U_{1\bar{Y}}(x)]\; dx,
\end{aligned}$$

avec $f(.)$ la densité de probabilité marginale du marqueur X, et où le résultat vient de l'hypothèse de randomisation. Si on pose $U_{0Y}(x) = -C_Y(x)$, $U_{0\bar{Y}}(x) = 0$, $U_{1Y}(x) = -[C_Y(x) + C_{++}(x) + C_Z(x)]$ et $U_{1\bar{Y}}(x) = -C_Z(x)$ alors la fonction d'utilité s'exprime comme :

$$\begin{aligned}
U(c) &= -\int_c^{+\infty} \mathrm{P}(Y=1|X=x,Z=0)\times C_Y(x)\; dF(x) \\
&\quad - \int_{-\infty}^{c} \mathrm{P}(Y=1|X=x,Z=1)\times [C_Y(x)+C_{++}(x)] + C_Z(x)\; dF(x) \\
&= \int_c^{+\infty} \mathrm{P}(Y=1|X=x,Z=1)\times [C_Y(x)+C_{++}(x)] - \mathrm{P}(Y=1|X=x,Z=0)\times C_Y(x) \\
&\quad + C_Z(x)\; dF(x) - \int_{-\infty}^{\infty} \mathrm{P}(Y=1|X=x,Z=1)\times [C_Y(x)+C_{++}(x)] + C_Z(x)\; dF(x)\,.
\end{aligned}$$

La deuxième intégrale étant une constante additive au regard de c, il est possible d'écrire que

$$\begin{aligned}
U(c) &\propto \int_c^{+\infty} \delta_w(x) + C_Z(x)\; dF(x) - C_M \\
&\propto \int_{-\infty}^{+\infty} \mathbb{1}(x>c)\times [\delta_w(x) + C_Z(x)]\; dF(x) - C_M \\
&\propto \Theta\,.
\end{aligned}$$

La seule différence avec les approches présentées précédemment se trouve donc dans la définition de la règle d'allocation qui est définie dans le cadre de cette thèse sur l'échelle du marqueur. Du fait de l'hypothèse de monotonicité de la différence de risques en fonction du marqueur formulée par Janes et al. (2014b), la règle d'allocation optimale présentée dans l'équation (3.3) est en réalité équivalente à la règle d'allocation optimale, notée $h^*(X)$, exprimée sur l'échelle du

marqueur :

$$h^*(X) = \mathbb{1}(X \leq c^*),$$

où c^* est le seuil optimal de marqueur défini dans l'article de Blangero et al. (2019). Comme la plupart des méthodes basées sur l'estimation de $\delta(X)$, ou $\delta_w(X)$, font l'hypothèse que ces quantités sont strictement monotones en fonction de X, cela signifie que le seuil qui maximise Θ est également le seuil qui maximise $U(.)$ (cela est également présenté dans l'article de Blangero et al. (2019)). Ainsi, les deux approches conduisent au même résultat, mais l'une est basée sur la valeur du marqueur, qui sera la règle concrètement utilisée en pratique clinique, et l'autre sur une valeur de différence de risque.

3.3 Perspectives

3.3.1 Amélioration de la précision de la méthode d'estimation

L'approche proposée précédemment dans l'article de Blangero et al. (2019) propose d'inclure la contrainte de randomisation liée à l'essai clinique dans le processus d'estimation du seuil optimal en définissant la distribution de marqueur dans un groupe de patients en fonction des distributions de marqueur dans les trois autres groupes, ainsi que des risques d'évènement moyens dans chaque bras de traitement :

$$F(x)^{(0Y)} = \left[F(x)^{(1Y)}\rho_1 + F(x)^{(1\bar{Y})}(1-\rho_1) - F(x)^{(0\bar{Y})}(1-\rho_0)\right]/\rho_0 \qquad \forall x, \tag{3.4}$$

où $F(.)^{(zl)}$ est la fonction de répartition du marqueur évaluée dans le groupe de patients zl. Ceci garantit que la contrainte de randomisation est respectée dans le processus d'estimation. En procédant ainsi, les données du quatrième groupe de patients ne sont pas utilisées dans le processus d'estimation, conduisant à une perte d'efficience. Une alternative à cette méthode serait d'inclure la contrainte de randomisation (3.4) directement dans l'algorithme d'échantillonnage des paramètres, en utilisant alors les données de l'ensemble des patients. L'objectif est donc ici de se servir également des données du quatrième groupe de patients afin d'améliorer encore les performances en matière de précision de la méthode.

En notant $\boldsymbol{\theta}$ le vecteur, de longueur d, des paramètres des lois de distribution théoriques retenues pour modéliser les distributions de marqueur dans chaque groupe de patients, ainsi que les risques moyens de survenue d'évènement dans chaque bras, et en notant de manière générique $p(.)$ une densité de probabilité, alors la distribution *a posteriori* de $\boldsymbol{\theta}$ peut s'exprimer comme suit :

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}^{(1\bar{Y})},\boldsymbol{x}^{(1Y)},\boldsymbol{x}^{(0\bar{Y})},\boldsymbol{x}^{(0Y)},\boldsymbol{y}^{(0)},\boldsymbol{y}^{(1)}) \propto\ & p(\boldsymbol{x}^{(1\bar{Y})}|\boldsymbol{\theta}_{1\bar{Y}})p(\boldsymbol{\theta}_{1\bar{Y}}) \times p(\boldsymbol{x}^{(1Y)}|\boldsymbol{\theta}_{1Y})p(\boldsymbol{\theta}_{1Y}) \\ & \times p(\boldsymbol{x}^{(0\bar{Y})}|\boldsymbol{\theta}_{0\bar{Y}})p(\boldsymbol{\theta}_{0\bar{Y}}) \times p(\boldsymbol{x}^{(0Y)}|\boldsymbol{\theta}) \\ & \times p(\boldsymbol{y}^{(0)}|\rho_0)p(\rho_0) \times p(\boldsymbol{y}^{(1)}|\rho_1)p(\rho_1), \end{aligned} \tag{3.5}$$

avec $\boldsymbol{x}^{(zl)}$ le vecteur des valeurs de marqueur observées dans l'échantillon de patients zl, et $\boldsymbol{y}^{(l)}$ le vecteur des valeurs de critère de jugement observées dans le bras de traitement z.

Dans cette expression, il a été considéré que les distributions des trois premiers groupes ($1\bar{Y}$, $1Y$, $0\bar{Y}$) sont connues explicitement, et de paramètres $\boldsymbol{\theta}_{zl}$. La distribution du quatrième groupe $0Y$ est exprimée en fonction des distributions fixées pour les trois autres groupes grâce à la contrainte de randomisation (3.4). Dans l'équation (3.5), les valeurs de marqueurs des quatre groupes de patients sont utilisées pour la distribution *a posteriori* des paramètres. Dans cette distribution *a posteriori*, un même paramètre intervient à deux endroits, par exemple le vecteur de paramètres $\boldsymbol{\theta}_{1\bar{Y}}$ intervient dans l'évaluation de la densité des valeurs du marqueur pour le groupe $1\bar{Y}$, mais également dans celle du groupe $0Y$, mais cette fois-ci sous une forme très complexe en raison de la forme de l'expression de la contrainte de randomisation (3.4). Ainsi, les distributions *a posteriori* conditionnelles des paramètres ne sont pas connues de façon explicite, et il n'est pas possible de recourir à un algorithme de type « Gibbs sampling » pour l'inférence. L'utilisation d'un algorithme de type Metropolis est une solution pour échantillonner des valeurs de paramètres. L'algorithme s'écrit de la manière suivante :

1. Définir des valeurs de départ $\boldsymbol{\theta}_0$.
2. Pour $t = 1, 2, \ldots$
   (a) Proposer une nouvelle valeur $\boldsymbol{\theta}_*$ à partir d'une distribution de proposition à l'itération t, $J_t(\boldsymbol{\theta}_* | \boldsymbol{\theta}_{t-1})$.
   (b) Calculer le ratio des densités,
   $$r = \frac{p(\boldsymbol{\theta}_* | \boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{\theta}_{t-1} | \boldsymbol{x}, \boldsymbol{y})}.$$
   (c) Définir
   $$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_* & \text{avec comme probabilité } \min(r, 1) \\ \boldsymbol{\theta}_{t-1} & \text{sinon} \end{cases}$$

Malgré son apparente simplicité, cet algorithme pose quelques problèmes techniques, notamment pour le choix de la distribution de proposition en cas de distribution multivariée. Le choix de cette distribution est crucial pour faciliter la convergence de l'algorithme vers la distribution *a posteriori* de $\boldsymbol{\theta}$.

On souhaite notamment que la distribution de proposition $J_t(.)$ permette d'échantillonner efficacement des valeurs de $\boldsymbol{\theta}$, en tenant compte de la corrélation entre les paramètres induite par l'équation de la randomisation et l'utilisation du quatrième groupe de patients. Cette corrélation est prise en compte au travers de la matrice de variance-covariance caractérisant la distribution $J_t(.)$, or cette matrice est difficile à fixer *a priori* car la corrélation entre les paramètres est difficilement quantifiable.

Une méthode pour approximer cette matrice de variance-covariance est donc d'avoir recours à un Metropolis adaptatif (Brooks et al., 2011). Le Metropolis adaptatif est une modification du Metropolis permettant d'optimiser la matrice de variance-covariance des paramètres échantillonnés. Plusieurs auteurs ont proposé leur algorithme permettant d'optimiser cette matrice,

celui qui est présenté est le Metropolis adaptatif robuste proposé par Vihola (2012) permettant d'optimiser la matrice de variance-covariance dans le but d'atteindre un taux d'acceptation des valeurs proposées fixé *a priori*. L'algorithme Metropolis classique est donc modifié comme suit :

1. Définir des valeurs de départ $\boldsymbol{\theta}_0$.
2. Pour $t = 1, 2, ...$
   (a) Proposer une nouvelle valeur $\boldsymbol{\theta}_* = \boldsymbol{\theta}_{t-1} + \boldsymbol{S}_{t-1}\boldsymbol{U}_t$ avec $\boldsymbol{U}_t$ un vecteur aléatoire indépendant tiré dans une distribution de Student multivariée, et $\boldsymbol{S}_{t-1}$ la matrice triangulaire issue de la décomposition de Cholesky de la matrice de variance-covariance empirique $\text{Cov}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1})$.
   (b) Calculer le ratio des densités,
   $$r = \frac{p(\boldsymbol{\theta}_*|\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{\theta}_{t-1}|\boldsymbol{x}, \boldsymbol{y})}.$$
   (c) Définir
   $$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_* & \text{avec comme probabilité } \min(r, 1) \\ \boldsymbol{\theta}_{t-1} & \text{sinon} \end{cases}$$
   (d) Définir $\boldsymbol{M}_t = \boldsymbol{S}_{t-1}\left(\boldsymbol{I} + \eta_t(\alpha_t - \alpha_*)\frac{\boldsymbol{U}_t\boldsymbol{U}_t'}{||\boldsymbol{U}||^2}\right)\boldsymbol{S}_{t-1}'$, avec $\boldsymbol{I}$ la matrice identité de taille $d \times d$, $\eta_t = \min(1, d \times t^{-2/3})$, α_t la proportion de θ_* acceptés lors des étapes 1 à t de l'algorithme et α_* le taux d'acceptation moyen souhaité (typiquement égal 0.234 lorsque le nombre de paramètres à échantillonner est élevé d'après Gelman et al. (2013)).

La matrice de variance-covariance $\boldsymbol{M}_t$ de $J_t(.)$ est donc modifiée à chaque itération de l'algorithme. Celui-ci doit se poursuivre jusqu'à ce que l'estimation de la matrice $\boldsymbol{M}_t$ soit stable. En revanche, l'ensemble des valeurs de $\boldsymbol{\theta}$ échantillonnées durant cette phase adaptative ne peut pas être utilisée comme étant un échantillon de la distribution *a posteriori* de $\boldsymbol{\theta}$ car la chaîne MCMC issue de la phase adaptative perd sa propriété d'ergodicité (Brooks et al., 2011). La phase adaptative doit donc être restreinte à l'étape de burn-in de l'algorithme MCMC. Une fois que $\boldsymbol{M}_t$ est stable, on stoppe la phase adaptative et on fixe $\boldsymbol{M}_t$ en réalisant un Metropolis classique pour échantillonner les valeurs de $\boldsymbol{\theta}$.

L'intérêt de cet approche est donc d'apporter un gain dans la précision de la méthode d'inférence présentée dans l'article en utilisant les données des quatre groupes de patients. Cependant, les propriétés de cette méthode doivent être démontrées dans le contexte de l'estimation du seuil optimal d'un marqueur prédictif par simulation.

3.3.2 Estimation des utilités moyennes

Dans l'article de Janes et al. (2014b), les auteurs fixent les valeurs de coûts moyens (présentés dans le Tableau 3.2) selon leur jugement personnel et selon les informations disponibles dans

la littérature concernant la maladie étudiée (le cancer du sein dans leur article), ainsi que sur les traitements étudiés. Il est également possible de collecter des données dans le cadre d'un essai clinique mesurant les coûts individuels liés aux traitements, à l'évènement d'intérêt, ainsi qu'à la mesure du marqueur. Les auteurs proposent alors de modéliser les coûts moyens pour obtenir l'estimation des coûts utilisés dans les formules du bénéfice moyen.

La même approche pourrait être envisagée pour modéliser les utilités définies dans la fonction d'utilité (3.1). Si on définit U l'utilité individuelle mesurée dans le cadre d'un essai clinique, alors il est possible d'estimer les utilités moyennes à l'aide du modèle suivant :

$$g[\mathbf{E}(U|Z,Y,X)] = \beta_0 + \beta_X X + \beta_Z Z + \beta_Y Y + \beta_{ZY} ZY + \beta_{ZX} ZX + \beta_{YX} YX + \beta_{ZYX} ZYX\,, \quad (3.6)$$

où $g(.)$ est la fonction de lien. Ce modèle diffère de celui proposé par Janes et al. (2014b) car il inclut les effets propres de Y, Z, ainsi que l'interaction entre Y et Z. Il est alors possible de calculer :

$$\begin{aligned}
U_{0\bar{Y}}(x) &= \psi(\beta_0 + \beta_X \times x)\,,\\
U_{0Y}(x) &= \psi[\beta_0 + \beta_Y + (\beta_X + \beta_{YX}) \times x]\,,\\
U_{1\bar{Y}}(x) &= \psi[\beta_0 + \beta_Z + (\beta_X + \beta_{ZX}) \times x]\,,\\
U_{1Y}(x) &= \psi[\beta_0 + \beta_Z + \beta_Y + \beta_{ZY} + (\beta_X + \beta_{ZX} + \beta_{YX} + \beta_{ZYX}) \times x]\,,
\end{aligned}$$

où $\psi(.)$ est la fonction de lien inverse. Si les utilités ne sont pas autorisées à dépendre du marqueur X, alors le modèle est simplifié et exprimé comme :

$$g[\mathbf{E}(U|Z,Y)] = \beta_0 + \beta_Z Z + \beta_Y Y + \beta_{ZY} ZY\,,$$

avec $U_{0\bar{Y}} = \psi(\beta_0)$, $U_{1\bar{Y}} = \psi(\beta_0 + \beta_Z)$, $U_{0Y} = \psi(\beta_0 + \beta_Y)$, et $U_{1Y} = \psi(\beta_0 + \beta_Z + \beta_Y + \beta_{ZY})$.

L'estimation des différentes utilités à l'aide de ces modèles permet d'introduire dans l'équation générale définie dans l'équation (3.1) des utilités estimées, avec leur incertitude, et non plus fixées de manière absolue, et pouvant dépendre éventuellement de covariables. Ces modèles peuvent également être utilisés pour généraliser la fonction d'utilité à l'évaluation de critères de jugement continus (la variable Y introduite dans le modèle prédisant les coûts ne sera alors plus binaire, mais continue). L'expression de la fonction d'utilité (3.1) est modifiée et s'exprime alors comme :

$$\begin{aligned}
U(c) = &\int_c^{+\infty} \int_{-\infty}^{\infty} \mathrm{P}(Y = y, X = x|Z = 0) \times U_0(x,y)\; dydx\\
&+ \int_{-\infty}^{c} \int_{-\infty}^{\infty} \mathrm{P}(Y = y, X = x|Z = 1) \times U_1(x,y)\; dydx\,,
\end{aligned}$$

où $U_z(x,y)$ est l'utilité prédite pour une valeur de marqueur $X = x$, une valeur du critère de jugement $Y = y$ dans le bras de traitement $Z = z$. Le modèle (3.6) peut alors être utilisé pour estimer $U_0(x,y) = \psi(\beta_0 + \beta_X \times x + \beta_Y \times y + \beta_{YX} \times yx)$ et $U_1(x,y) = \psi[\beta_0 + \beta_Z + (\beta_X + \beta_{ZX}) \times x + (\beta_Y + \beta_{ZY}) \times y + (\beta_{YX} + \beta_{ZYX}) \times yx]$. On pourrait par exemple envisager ce cas de figure

en cardiologie, pour le choix d'un traitement antihypertenseur. En effet, la méthode présentée jusqu'à présent nécessite la définition d'un critère binaire de succès du traitement. Il faut donc fixer un seuil arbitraire de variation de pression artérielle à partir duquel le patient est considéré comme répondeur au traitement. La méthode présentée juste au dessus permettrait de considérer comme critère de jugement la valeur de pression artérielle en elle-même, s'affranchissant ainsi du choix d'un seuil arbitraire pour binariser le critère de jugement.

Si les utilités sont estimées à l'aide de l'un de ces deux modèles, il est nécessaire de tenir compte de l'incertitude de leur estimation lorsqu'elles sont utilisées dans la fonction d'utilité. Janes et al. (2014b) proposent d'utiliser des méthodes bootstrap afin de tenir compte de l'incertitude dans l'estimation des utilités, ainsi il est nécessaire de réestimer les utilités pour chaque échantillon bootstrap. Cependant la méthode d'estimation présentée précédemment dans l'article (Blangero et al., 2019) peut être adaptée pour intégrer l'estimation des paramètres du modèle (3.6) dans le processus d'inférence Bayésienne et tenir compte de l'incertitude dans leur estimation.

Néanmoins, ces extensions nécessitent de pouvoir définir correctement les modèles d'utilité, avec des formes fonctionnelles vraisemblablement complexes et de multiples interactions. La fiabilité du seuil optimal fourni dépend alors fortement de la fiabilité des modèles d'utilité, et du choix de la métrique d'utilité, comme la quantité de vie ou la qualité de vie. Ceci ne peut donc pas être envisagé systématiquement.

Dantan et al. (2018) ont également proposé une fonction d'utilité permettant d'estimer le seuil optimal d'un marqueur prédictif pour un critère de jugement de type « temps jusqu'à évènement ». En faisant l'hypothèse que les utilités sont constantes au cours du temps, il est possible d'exprimer la fonction d'utilité sous la forme :

$$\begin{aligned} U_t(c) = [1-F(c)]\{U_{0\bar{Y}}\mathbf{E}[\min(T,t)|X>c,Z=0] + U_{0Y}(t-\mathbf{E}[\min(T,t)|X>c,Z=0])\} \\ + F(c)\{U_{1\bar{Y}}\mathbf{E}[\min(T,t)|X\leq c,Z=1] + U_{1Y}(t-\mathbf{E}[\min(T,t)|X\leq c,Z=1])\} \end{aligned}$$

avec t le délai d'observation fixé, T le temps jusqu'à la survenue de l'évènement étudié et $\mathbf{E}[\min(T,t)|g,Z=z]$ la moyenne de survie restreinte au temps t dans le bras $Z=z$ pour des patients ayant des valeurs de marqueur $g \in \{X>c, X\leq c\}$.

Dans leur article, les auteurs proposent une approche non paramétrique d'estimation du seuil optimal en s'appuyant sur la fonction de répartition empirique du marqueur ainsi que sur les courbes de survie estimées par la méthode de Kaplan-Meier. Cette approche implique une grande incertitude dans l'estimation du seuil, et les auteurs ne présentent pas les résultats liés à la probabilité de couverture du seuil optimal obtenus par l'intervalle de confiance bootstrap dans leurs simulations. Cependant, l'approche Bayésienne présentée précédemment pourrait bien être adaptée pour ajuster des distributions théoriques de marqueur et des distributions théoriques de survie pour estimer le seuil optimal de manière plus précise.

3.3.3 Développement d'un package R

Un package `R` codé en S4 est en cours de développement et devrait prochainement être soumis au CRAN (Comprehensive R Archive Network) afin de promouvoir la méthode d'estimation du seuil optimal décrite précédemment.

L'objectif de ce package est de mettre à la disposition des utilisateurs un outil facile à utiliser afin d'estimer le seuil optimal d'un marqueur ainsi que son intervalle de confiance à l'aide la méthode présentée précédemment (Blangero et al., 2019). Les utilisateurs pourront s'ils le souhaitent :

- Définir eux-mêmes des distributions théoriques de marqueur en fournissant la fonction de densité de cette distribution, la fonction de répartition, ainsi que la dérivée de la fonction de densité
- Fournir directement les chaînes MCMC des paramètres de chaque distribution qu'ils auront obtenus par eux-mêmes, afin d'obtenir la chaîne MCMC du seuil optimal
- Utiliser des distributions théoriques plus communes et déjà intégrées dans le package (loi normale, loi log-normale, loi de Student, loi gamma, loi logistique)
- Utiliser un algorithme de Metropolis adaptatif selon la méthode définie dans la sous-section 3.3.1
- Afficher des outils graphiques de diagnostic du processus MCMC, ou bien de faire des graphiques permettant d'évaluer l'impact du marqueur (ex : courbes de décision)
- Obtenir un ensemble d'indicateurs résumant les performances du marqueur

En reprenant l'exemple du marqueur génétique *DDR2* issu de l'essai PETACC-8, il est possible d'illustrer les fonctionnalités du package qui ont pour le moment été développées. La première étape consiste à stocker les valeurs de marqueur des quatre groupes de patients dans quatre objets `R` différents. En supposant que le jeu de données est stocké dans l'objet `data`, que la variable mesurant le critère de jugement est nommée `Y`, et que la variable mesurant le traitement administré aux patients est nommée `trt` alors on écrit :

```
> DDR20Y <- data$DDR2[data$Y==1 & data$trt==0,]
> DDR20Yb <- data$DDR2[data$Y==0 & data$trt==0,]
> DDR21Y <- data$DDR2[data$Y==1 & data$trt==1,]
> DDR21Yb <- data$DDR2[data$Y==0 & data$trt==1,]
```

Dans l'article de Blangero et al. (2019), il a été décidé d'exprimer la distribution du marqueur du groupe de patients $0Y$ en fonction des distributions de marqueur des trois autres groupes. Pour les trois autres groupes de patients, les distributions de marqueur qui ont été retenues sont les suivantes :

- Groupe $0\bar{Y}$: une distribution Gaussienne
- Groupe $1Y$: une distribution de Student
- Groupe $1\bar{Y}$: une distribution log-normale

Sous R, il est alors possible de créer des objets qui vont ajuster ces distributions aux valeurs de marqueur :

```
> fit0Y <- undefined(DDR20Y)
> fit0Yb <- fitNormalDist(DDR2Yb)
> fit1Y <- fitStudentDist(DDR21Y,ini=list(list(mu=0.04,sd=0.01,df=5),
+ list(mu=0.07,sd=0.015,df=6),list(mu=0.12,sd=0.02,df=7)),thin=5,
+ burnin=2000)
> fit1Yb <- fitLogNormalDist(DDR21Yb)
```

Pour l'objet `fit0Y` la fonction `undefined` est utilisée pour indiquer au package que la distribution de ce marqueur sera exprimée en fonction des trois autres. Pour les objets `fit0Yb` et `fit1Yb` les lois de probabilité sont simples et il n'y a pas besoin d'algorithme MCMC pour obtenir la distribution *a posteriori* des paramètres. En revanche, pour l'objet `fit1Y`, la distribution *a posteriori* de ses paramètres n'est pas connue de façon explicite, il est donc nécessaire d'avoir recours un algorithme MCMC (intégré dans la fonction d'estimation du seuil optimal). Pour cet exemple, trois chaînes MCMC sont calculées, avec des valeurs initiales différentes pour chacune d'entre elles afin d'évaluer la convergence de l'algorithme. Il est désormais possible d'utiliser la fonction principale du package, appelée `optThreshEst`, qui va permettre d'estimer le seuil optimal du marqueur *DDR2* :

```
> threshRes<-optThreshEst(EvtRefDist=fit0Y,
+ NoEvtRefDist=fit0Yb,
+ EvtInnovDist=fit1Y,
+ NoEvtInnovDist=fit1Yb,
+ lowRef = FALSE,
+ toxRef = FALSE,
+ r=0,
+ le.MCMC = 5000,
+ plot=TRUE)
```

Ici, l'option `lowRef` permet de spécifier à la fonction quel traitement est préféré pour des valeurs basses de marqueur. S'il s'agit du traitement de référence alors `lowRef = TRUE`, autrement `lowRef=FALSE`. L'option toxRef permet de spécifier à la fonction quel traitement est le plus toxique à efficacité comparable. S'il s'agit du traitement de référence alors `toxRef=TRUE` autrement `toxRef=FALSE`. L'option *r* permet de spécifier le ratio $\frac{C_Z}{C_Y}$, aussi appelé ratio *r* dans l'article. L'option `le.MCMC` permet de spécifier à la fonction la longueur de la chaîne MCMC désirée en sortie, et l'option `plot` permet à l'utilisateur d'afficher des diagnostics graphiques du processus MCMC s'il le souhaite.

L'exécution de cette fonction peut prendre un temps non négligeable selon la complexité des distributions ajustées sur les valeurs de marqueur et de la longueur de la chaîne MCMC souhaitée. Une fois l'échantillonnage des paramètres de distribution réalisé, la fonction doit encore optimiser la fonction d'utilité pour chaque itération de l'algorithme MCMC, ce qui peut également prendre du temps. Une option, nommée `progress.bar` a donc été rajoutée à la fonction pour connaître l'étape à laquelle se trouve la fonction `optThreshEst`.

Une fois que le processus est terminé, il est possible d'afficher un résumé des informations contenues dans l'objet `threshRes` grâce à la fonction `summary` :

```
> summary(threshRes)
Decision rule: The innovative treatment is preferred for low values
of the marker.

Median risk in the reference group: 0.16 [0.13; 0.19]
Median risk in the innovative group: 0.17 [0.13; 0.2]

Summary statistics of the marker:
    Min.   1st Qu.    Median   3rd Qu.      Max.
0.005917  0.059048  0.070106  0.082316  0.178485

Optimal threshold estimate (median):
[1]  0.0731033

Credible intervals of the optimal threshold
- Percentile method (95%)
[1]  0.06823982  0.07816284
- Highest Posterior Density (95%)
[1]  0.06798578  0.07788085

Median risk in the reference group under marker-based strategy:
0.12 [0.09; 0.16]
Median risk in the innovative group under marker-based strategy:
0.15 [0.11; 0.18]

Benefit of the marker-based strategy
- In the reference group: 0.04 [0.02; 0.07]
- In the innovative group: 0.02 [0; 0.05]

Percentage of NA values returned during the estimation process: 3.3%
```

Ce résumé rappelle la règle de décision qui a été utilisée dans l'optimisation de la fonction d'utilité et présente plusieurs indicateurs statistiques :

- Le risque médian dans chaque bras de traitement ainsi que son intervalle de crédibilité
- Les statistiques descriptives de base du marqueur (min, max, moyenne et quartiles)
- L'estimation ponctuelle du seuil optimale (médiane de la chaîne MCMC)
- Les intervalles de crédibilité du seuil optimal (par la méthode des percentiles et la méthode « Highest Posterior Density », HPD)
- La médiane du risque d'évènement dans chaque bras conditionnellement à la règle de décision spécifiée

- Le bénéfice moyen
- Le pourcentage d'itérations MCMC pour lesquelles il n'a pas été possible d'estimer le seuil optimal

Les estimations ponctuelles ainsi que les intervalle de crédibilité de ces indicateurs peuvent être extraits de ce résumé grâce aux fonction `estimates` et `credibleIntervals` :

```
> estimates(threshRes)
$optThresh
[1]  0.0731033

$risks
 Reference  Innovative
 0.1631386  0.1681877

$markerBasedRisks
 Reference  Innovative
0.12299554  0.14522362

$benefits
 Reference  Innovative
0.04014307  0.02296412

> credibleIntervals(threshRes)
                                        2.5%       97.5%
OptThresh CI                     0.0682398186  0.07816284
Reference risk CI                0.1282945539  0.18916464
Innovative risk CI               0.1331674616  0.19670942
Reference marker-based risk CI   0.0915107855  0.15731149
Innovative marker-based risk CI  0.1065257717  0.17643804
Reference benefit CI             0.0360630025  0.07317775
Innovative benefit CI            0.0006813446  0.05170964
```

Il est également possible d'afficher les intervalles de crédibilité par la méthode HPD :

```
> credibleIntervals(threshRes,hpd=TRUE)
                                       lower       upper
OptThresh CI                     0.067985776  0.07788085
Reference risk CI                0.127933889  0.18875390
Innovative risk CI               0.132237741  0.19528685
Reference marker-based risk CI   0.096084565  0.16094469
Innovative marker-based risk CI  0.105465538  0.17515778
Reference benefit CI             0.032114508  0.06821386
Innovative benefit CI            0.001974908  0.05993679
```

Ces fonctions permettent d'obtenir rapidement les informations contenues dans le résumé. Il

est également possible d'utiliser la fonction `plot` directement sur l'objet `threshRes` afin de visualiser la distribution du seuil optimal et de personnaliser ce graphique en utilisant les options classiques de cette fonction :

```
> plot(threshRes,main="Distribution du seuil optimal",
+ xlab="Seuil optimal",ylab="Densite",cex.lab=1.5,cex.main=1.8)
```

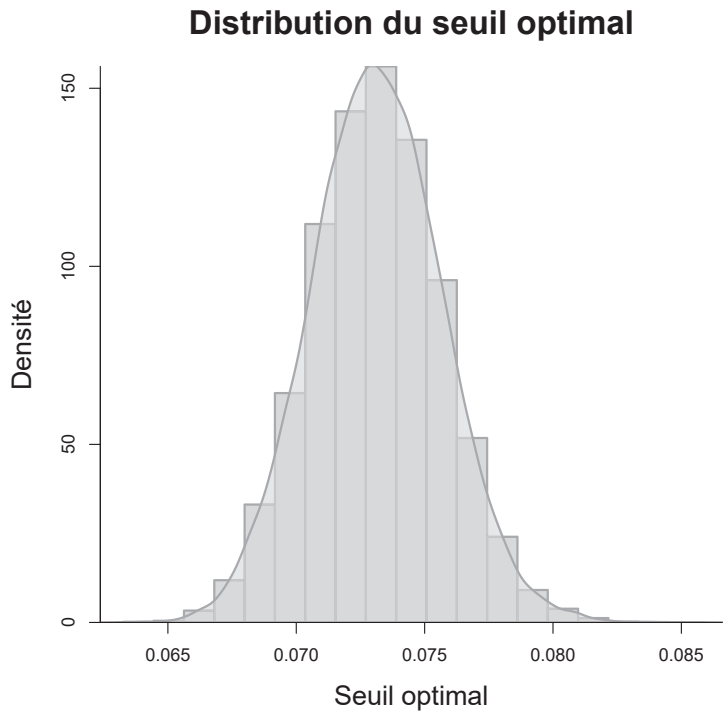


FIGURE 3.1 – Histogramme représentant la distribution du seuil optimal du gène *DDR2*

La Figure 3.1 présente la sortie graphique de cette fonction. Enfin, une dernière fonction est disponible et permet de visualiser les courbes de décision associées au marqueur en faisant varier le ratio $\frac{C_Z}{C_Y}$. Cette fonction peut mettre beaucoup de temps à s'exécuter car elle doit, pour chaque ratio donné en argument, optimiser de nouveau les fonctions d'utilité afin de calculer le seuil optimal du marqueur pour chacun d'entre eux, et produire un intervalle de confiance pour chaque point. Ainsi il est recommandé de stocker le résultat dans un objet `R` afin de ne pas refaire tous les calculs si l'utilisateur souhaite afficher de nouveau la courbe de décision à l'écran :

```
> ru<-decisionCurve(threshRes,r=seq(0,0.1,length.out=6))
> plot(ru,which=2,confband=FALSE,main2="Courbe de decision
+ (relative)",xlab="Ratio r",ylab2="Utilite relative",cex.lab=1.5,
+ cex.main=1.8)
```

La Figure 3.2 présente la sortie graphique de cette fonction. L'option `which` permet à l'utilisateur de visualiser soit les courbes de décision des stratégies extrêmes et de celle associée au marqueur, soit de visualiser la courbe de décision de l'utilité relative comme décrit dans l'article

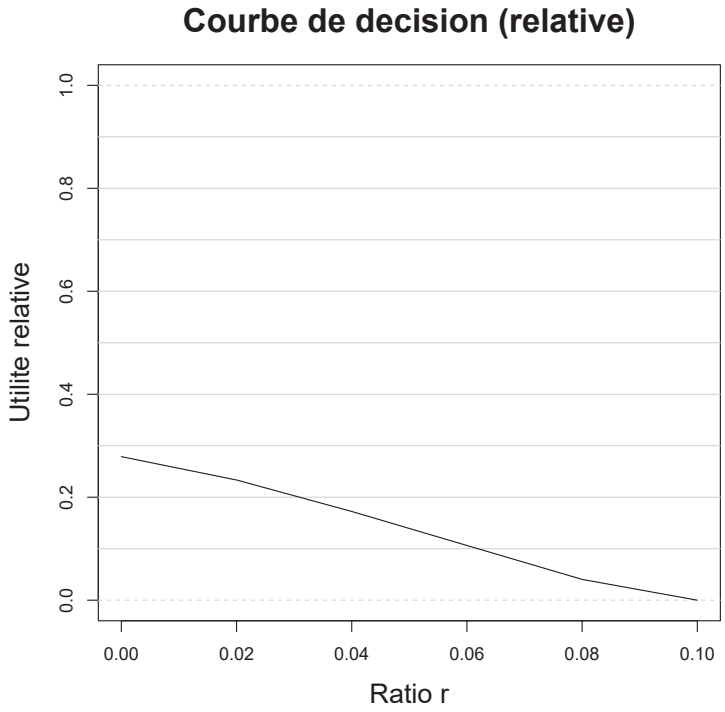


FIGURE 3.2 – Courbe de décision présentant l'évolution de l'utilité relative du gène *DDR2* en fonction du ratio $\frac{C_Z}{C_Y}$ (ou ratio r)

de Blangero et al. (2019). L'option `confband` permet à l'utilisateur de choisir s'il souhaite afficher la bande de confiance autour de la courbe de décision. Toutes ces fonctionnalités sont donc déjà implémentées dans le package mais elles ont besoin d'être validées après une série de tests avant que le package ne soit soumis au CRAN. De plus, il reste des développements à réaliser concernant l'intégration de l'algorithme MCMC adaptatif afin de proposer aux utilisateurs d'utiliser les quatre groupes de patients pour l'estimation du seuil optimal du marqueur.

3.4 Bilan du chapitre 3

L'objectif de ce chapitre était de présenter des méthodes permettant d'estimer le seuil optimal d'un marqueur utilisé pour guider le choix de traitement. La méthode proposée dans ce chapitre repose sur la définition d'une fonction d'utilité, à l'image de ce qui est déjà réalisé pour l'analyse des marqueurs diagnostiques et pronostiques.

Il a été démontré que la fonction d'utilité proposée est une expression plus générale du bénéfice moyen défini par Janes et al. (2014b). Il a également été démontré que selon les hypothèses réalisées sur les mesures d'utilités alors le seuil qui maximise la fonction d'utilité est également le seuil de marqueur défini pour plusieurs méthodes (Vickers et al., 2007; Janes et al., 2014a,b).

L'approche proposée fait l'hypothèse qu'il existe un seul seuil de marqueur à estimer ; c'est également le cas des approches décrites dans la littérature (Janes et al., 2014b,a). Des auteurs se sont intéressés à l'évaluation d'une stratégie de traitement basée sur un marqueur pour lequel

l'hypothèse de monotonicité de $\delta(X)$ en fonction de X n'est pas vérifiée (Matsouaka et al., 2014) ; cela signifie que sur l'échelle du marqueur il peut ne pas y avoir un seuil unique et que la règle de décision optimale ne peut plus être exprimée sous la forme de $h^*(X) = \mathbb{1}(X \leq c^*)$. Cependant ces approches reposent sur des modèles de prédiction complexes qui, s'ils ne sont pas corrects, peuvent amener à des solutions sous-optimales.

La méthode d'estimation de la fonction d'utilité repose sur l'ajustement de distributions théoriques de marqueurs pour quatre groupes de patients : ceux qui ont développé l'évènement d'intérêt ou non dans chacun des bras de traitement. Il s'agit là de la principale difficulté dans l'application de cette méthode, cependant l'alternative à cette approche serait de modéliser le risque de survenue d'évènement dans chaque bras, conditionnellement aux valeurs du marqueur comme proposé par Vickers et al. (2007), Janes et al. (2014a) et Janes et al. (2014b) ce qui peut parfois s'avérer difficile lorsque les hypothèses du modèle utilisé ne sont pas vérifiées. L'avantage de la méthode proposée réside dans le fait que des distributions très souples de marqueur peuvent être définies, et que le cadre de l'inférence Bayésienne permet de faciliter l'estimation du seuil optimal du marqueur.

Il a été montré par la suite que cette fonction d'utilité pouvait facilement être généralisée à l'analyse des critères de jugement continus dans le cas où les utilités individuelles sont récoltées dans le cadre de l'essai clinique.

Conclusion

Dans ce travail de thèse, un ensemble de méthodes a été présenté concernant l'évaluation et l'utilisation des marqueurs prédictifs quantitatifs. Ces méthodes vont de la quantification du caractère prédictif global du marqueur, à la détermination d'un seuil optimal pour son utilisation en pratique clinique et l'aide à la décision médicale.

Une revue de la littérature importante a été réalisée concernant les méthodes de quantification du caractère prédictif de ces marqueurs, permettant de les catégoriser en deux familles de méthodes : celles évaluant le différentiel de réponses aux deux traitements à l'échelle individuelle, et celles évaluant ce différentiel à l'échelle d'une population. Cette distinction a permis de mieux comprendre les contraintes liées à ces approches et leur intérêt. Bien que leur résultat soit plus facile à interpréter, les méthodes issues de la première famille reposent sur de nombreuses hypothèses difficiles à évaluer, et souvent non vérifiées. Un nouvel indicateur, s'inscrivant dans la deuxième famille de méthodes, a ainsi pu être proposé en étendant l'utilisation des courbes ROC à l'analyse des marqueurs prédictifs. Cette méthode correspond à une première étape dans l'identification de marqueurs pouvant avoir un intérêt dans le choix du traitement.

Concernant la détermination du seuil optimal du marqueur, une fonction d'utilité a été définie et sa connexion avec d'autres fonctions proposées dans la littérature a pu être démontrée. Une méthode d'inférence Bayésienne a également été proposée, ce choix peut notamment se justifier par la simplicité d'utilisation des algorithmes MCMC, et leur capacité à tenir compte de l'incertitude dans l'estimation de chaque paramètre ; des extensions sont envisagées afin de rendre cette méthode d'estimation plus efficace.

Toutes ces méthodes ont pu être appliquées aux données de l'essai clinique PETACC-8 dans le contexte du cancer colorectal. Ces analyses ont permis de mettre en évidence l'intérêt de l'utilisation du niveau d'amplification du gène *DDR2* pour choisir entre le FOLFOX4 et le FOLFOX4 combiné au cetuximab. L'utilisation de ce marqueur en pratique clinique doit encore être discutée, et ses performances doivent être validées au travers d'un essai évaluant spécifiquement une stratégie de traitement axée sur l'utilisation de ce marqueur.

De manière générale, il sera important de promouvoir cette méthode d'estimation de seuil optimal dans le monde de la recherche clinique afin de sensibiliser ses acteurs à l'intérêt de cette méthode. En effet, il est fréquent que des seuils de marqueur soient fixés de manière arbitraire ou bien sans tenir compte des coûts cliniques. Dans ce contexte, ce travail permettre peut-être de faire évoluer les pratiques et d'orienter la recherche vers de nouveaux développements méthodologiques dans ce domaine. A ce titre, un package `R` est en cours de finalisation afin de

promouvoir ces méthodes.

Bibliographie

André, T., Boni, C., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Bonetti, A., Clingan, P., Bridgewater, J., Rivera, F., and de Gramont, A. (2009). Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *Journal of Clinical Oncology*, 27(19) :3109–3116.

Biomarker Definitions Working Group (2001). Biomarkers and surrogate endpoints : Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, 69(3) :89–95.

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013a). 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.

Blanche, P., Dartigues, J. F., and Jacqmin-Gadda, H. (2013b). Review and comparison of roc curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5) :687–704.

Blangero, Y., Rabilloud, M., Ecochard, R., and Subtil, F. (2019). A bayesian method to estimate the optimal threshold of a marker used to select patients' treatment. *Statistical Methods in Medical Research*.

Bokemeyer, C., Bondarenko, I., Hartmann, J. T., de Braud, F., Schuch, G., Zubel, A., Celik, I., M., S., and Koralewski, P. (2011). Efficacy according to biomarker status of cetuximab plus FOLFOX-4 as first-line treatment for metastatic colorectal cancer : the OPUS study. *Annals of Oncology*, 22(7) :2535–2546.

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton.

Byar, D. P. (1985). Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine*, 4(3) :255–263.

Carlomagno, N., Incollingo, P., Tammaro, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M., Minieri, G., Pisani, A., Riccio, E., Sabbatini, M., Bracale, U., Calogero, A., Dodaro, C., and Santangelo, M. (2017). Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millenium : a breakthrough in gastric cancer. *BioMed research international*.

Cohen, M., Williams, G., Johnson, J., Duan, J., Gobburu, J., Rahman, A., Benson, K., Leighton, J., Kim, S., Wood, R., Rothmann, M., Chen, G., U, K., Staten, A., and Pazdur, R. (2002). Approval summary for imatinib mesylate capsules in the treatment of chronic myelogenous leukemia. *Clinical Cancer Research*, 8(5) :935–942.

Dantan, E., Foucher, Y., Lorent, M., Giral, M., and Tessier, P. (2018). Optimal threshold estimator of a prognostic marker by maximizing a time-dependent expected utility function for a patient-centered stratified medicine. *Statistical Methods in Medical Research*, 27(6) :1847–1859.

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

De Roock, W., Piessevaux, H., de Schutter, J., Janssens, M., De Hertog, G., Personeni, N., Biesmans, B., Van Laethem, J., Peeters, M., Humblet, Y., Van Cutsem, E., and Tejpar, S. (2008). KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. *Annals of Oncology*, 19(3) :508–515.

Defossez, G., Le Guyader-Peyroud, S., Uhry, Z., Grosclaude, P., Remontet, L., Colonna, M., Dantony, E., Delafosse, P., Molinié, F., Woronoff, A., Bouvier, A., Bossard, N., and Monnereau, A. (2019). *Estimation nationale de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Etude à partir des registres des cancers du réseau Francim - Résultats préliminaires*. Santé Publique France.

DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves : A nonparametric approach. *Biometrics*, 44(3) :837–845.

Di Fiore, F., Blanchard, F., Charbonnier, F., Le Pessot, F., Lamy, A., Galais, M., Bastit, L., Killian, A., Sesboüé, R., Tuech, J.-J., Queuniet, A., Paillot, B., Sabourin, J., Michot, F., Michel, P., and Frebourg, T. (2007). Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by cetuximab plus chemotherapy. *British Journal of Cancer*, 96(8) :1166–1169.

Dodd, L. and Pepe, M. S. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98(462) :409–417.

Douillard, J.-Y., Rong, A., and Sidhu, R. (2013). RAS mutations in colorectal cancer. *New England Journal of Medicine*, 369(22) :2159–2160.

Douillard, J.-Y., Siena, S., Cassidy, J., Tabernero, J., Burkes, R., Barugel, M., Humblet, Y., Bodoky, G., Cunningham, D., Jassem, J., F., R., Kocákova, I., Ruff, P., Błasińska-Morawiec, M.,

Šmakal, M., Canon, J.-L., Rother, M., Oliner, K. S., Wolf, M., and Gansert, J. (2010). Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer : the PRIME study. *Journal of Clinical Oncology*, 28(31) :4697–4705.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. CRC Press, Boca Raton.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*. CRC Press, Boca Raton.

Gerszten, R. and Wang, T. (2008). The search of new cardiovascular biomarkers. *Nature*, 451(7181) :949–952.

Gill, S., Loprinzi, C. L., Sargent, D. J., Thomé, S. D., Alberts, S. R., Haller, D. G., Benedetti, J., Francini, G., Shepherd, L. E., Seitz, J. F., Labianca, R., Chen, W., Cha, S. S., Heldebrant, M. P., and Goldberg, R. M. (2004). Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer : Who benefits and by how much ? *Journal of Clinical Oncology*, 22(10) :1797–1806.

Hanley, J. and McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1) :29–36.

Hernan, M. and Robins, J. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7) :578–586.

Hochhaus, A., Larson, R., Guilhot, F., Radich, J., Branford, S., Hughes, T., Baccarani, M., Deininger, M., Cervantes, F., Fujihara, S., Ortmann, C., Menssen, H., Kantarjian, H., O'Brien, S., Druker, B., and IRIS Investigators (2017). Long-term outcomes of imatinib treatment for chronic myeloid leukemia. *The New England journal of medicine*, 376(10) :917–927.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3) :293–325.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396) :945–960.

Huang, Y., Gilbert, P. B., and Janes, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. *Biometrics*, 68(3) :687–696.

Huang, Y., Laber, E., and Janes, H. (2015). Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics*, 16(2) :383–399.

Hung, H. and Chiang, C.-T. (2010). Estimation methods for time-dependent AUC models with survival data. *The Canadian Journal of Statistics*, 38(1) :8–26.

Janes, H., Brown, M., and Pepe, M. (2015a). Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in Medicine*, 34(27) :3503–3515.

Janes, H., Brown, M. D., Pepe, M. S., and Huang, Y. (2014a). An approach to evaluating and comparing biomarkers for patient treatment selection. *International Journal of Biostatistics*, 10(1) :99–121.

Janes, H., Pepe, M., McShane, L., Sargent, D., and Heagerty, P. (2015b). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute*, 107(8).

Janes, H., Pepe, M. S., Bossuyt, P. M., and Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*, 154(4) :253–259.

Janes, H., Pepe, M. S., and Huang, Y. (2014b). A framework for evaluating markers used to select patient treatment. *Medical Decision Making*, 34(2) :159–167.

Jund, J., Rabilloud, M., Wallon, M., and Ecochard, R. (2005). Methods to estimate the optimal threshold for normally or log-normally distributed biological tests. *Medical Decision Making*, 25(4) :406–415.

Kuebler, J. P., Wieand, H. S., O'Connell, M. J., Smith, R. E., Colangelo, L. H., Yothers, G., Petrelli, N. J., Findlay, M. P., Seay, T. E., Atkins, J. N., Zapas, John, L., Wendall Goodwin, J., Fehrenbacher, L., Ramanathan, R. K., Conley, B. A., Flynn, P. J., Soori, G., Colman, L. K., Levine, E. A., Lanier, K. S., and Wolmark, N. (2007). Oxaliplatin combined with weekly bolus fluorouracil and leucovorin as surgical adjuvant chemotherapy for stage II and III colon cancer : Results from NSABP C-07. *Journal of Clinical Oncology*, 25(16) :2198–2204.

Li, L., Greene, T., and Hu, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, 27(8) :2264–2278.

Lièvre, A., Bachet, J. B., Boige, V., Cayre, A., Le Corre, D., Buc, E., Ychou, M., Bouché, O., Landi, B., Louvet, C., André, T., Bibeau, F., Diebold, M., Rougier, P., Ducreux, M., Tomasic, G., Emile, J., Penault-Llorca, F., and Laurent-Puig, P. (2008). KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. *Journal of Clinical Oncology*, 26(3) :374–379.

Mamounas, E., Wieand, S., Wolmark, N., Bear, H. D., Atkins, J. N., Song, K., Jones, J., and Rockette, H. (1999). Comparative efficacy of adjuvant chemotherapy in patients with Dukes' B versus Dukes' C colon cancer : Results from four national surgical adjuvant breast and bowel project adjuvant studies (C-01, C-02, C-03, and C-04). *Journal of Clinical Oncology*, 17(5) :1349–1355.

Matsouaka, R., Li, J., and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics*, 70(3) :489–499.

McClish, D. (1989). Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3) :190–195.

Nedelman, J., Heagerty, P., and Lawrence, C. (1992). Quantitative pcr : Procedures and precisions. *Bulletin of Mathematical Biology*, 54(4) :477–502.

Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *The Computer Journal*, 7(4) :308–313.

O'Brien, S., Guilhot, F., Larson, R., Gathmann, I., Baccarani, M., Cervantes, F., Cornelissen, J., Fischer, T., Hochhaus, A., Hugues, T., Lechner, K., Nielsen, J., Rousselot, P., Reiffers, J., Saglio, G., Shepherd, J., Simonsson, B., Gratwohl, A., Goldman, J., Kantarjian, H., Taylor, K., Verhoef, G., Bolton, A., Capdeville, R., and Druker, BJ, I. (2003). Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *The New England journal of medicine*, 48(11) :994–1004.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Ransohoff, D. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nature*, 4(4) :309–314.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427) :846–866.

Rubin, D. (2005). Causal inference using potential outcomes : Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469) :322–331.

Sargent, D. J., Conley, B. A., Allegra, C., and Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9) :2020–2027.

Schiffer, E. (2009). The 2nd anual oncology biomarkers conference. *Biomarkers in Medicine*, 3(2) :203–209.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4) :874–883.

Song, X. and Zhou, X. (2011). Evaluating markers for treatment selection based on survival time. *Statistics in Medicine*, 30(18) :2251–2264.

Sox, H., Higgins, M., and Owens, D. (2013). *Medical Decision Making*. Wiley-Blackwell, Chichester.

Subtil, F. and Rabilloud, M. (2015). An enhancement of roc curves made them clinically relevant for diagnostic-test comparison and optimal threshold determination. *Journal of Clinical Epidemiology*, 68(7) :752–759.

Taieb, J., Tabernero, J., Mini, E., Subtil, F., Folprecht, G., Van Laethem, J., Thaler, J., Bridgewater, J., Petersen, L., Blons, H., Collette, L., Van Cutsem, E., Rougier, P., Salazar, R., Bedenne, L., Emile, J., Laurent-Puig, P., and Lepage, C. (2014). Oxaliplatin, fluorouracil, and leucovorin with or without cetuximab in patients with resected stage iii colon cancer (petacc-8) : an open-label, randomised phase 3 trial. *The Lancet Oncology*, 15(8) :862–873.

Tajik, P., Zwinderman, A., Mol, B., and Bossuyt, P. (2013). Trial designs for personalizing cancer care : a systematic review and classification. *Clinical Cancer Research*, 19(17) :4578–4588.

Viallon, V. and Latouche, A. (2011). Discrimination measures for survival outcomes : Connection between the AUC and the predictiveness curve. *Biometrical Journal*, 53(2) :217–236.

Vickers, A., Kattan, M., and Sargent, D. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, 8 :14.

Vihola, M. (2012). Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statistics and computing*, 22(5) :997–1008.

Wong, M. and Pollock, C. (2014). Biomarkers in kidney fibrosis : are they useful ? *Kidney international supplements*, 4(1) :79–83.

Yothers, G., O'Connell, M. J., Allegra, C. J., Kuebler, J. P., Colangelo, L. H., Petrelli, N. J., and Wolmark, N. (2011). Oxaliplatin as adjuvant therapy for colon cancer : updated results of NSABP C-07 trial, including survival and subset analyses. *Journal of Clinical Oncology*, 29(28) :3768–3774.

Zhang, Z., Ma, S. F., Nie, L., and Soon, G. (2017). A quantitative concordance measure for comparing and combining treatment selection markers. *The International Journal of Biostatistics*, 13(1).

Zhang, Z., Nie, L., Soon, G., and Liu, A. (2014). The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. *The Annals of Applied Statistics*, 8(4) :2336–2355.

Annexes

Annexe A

Calcul de la valeur maximale du gain total en fonction des risques moyens dans chaque bras

Cette annexe présente la démonstration permettant de calculer la valeur maximale atteignable par le gain total (TG) en fonction des risques moyens dans chaque bras de traitement (résultats présentés dans la sous-section 2.4.1). Pour rappel la formule du TG est la suivante :

$$\mathrm{TG} = \int |\delta(x) - (\rho_1 - \rho_0)| \, dF_\delta \, .$$

Il est possible de calculer la valeur maximale du TG de manière graphique en utilisant les courbes de risque associées au marqueur parfait. Si l'on reprend la définition donnée d'un marqueur parfait dans la sous-section 2.1.2, en retenant l'approche la plus conservatrice, alors soit q_2 soit q_3 devait être nul.

En prenant le cas où $q_2 = 0$, alors $q_1 = \rho_0$, $q_4 = \rho_1$ et $q_3 = 1 - \rho_0 - \rho_1$. La Figure A.1 présente un exemple de courbes de risque associées à un marqueur parfait pour lequel $q_2 = 0$.

Dans le cas du marqueur présenté dans la Figure A.1, le TG s'exprime de la manière suivante :

$$\begin{aligned}
\mathrm{TG} &= \int_0^{q_4} |1 + \rho_0 - \rho_1| \, dF_\delta + \int_{q_4}^{1-q_1} |\rho_0 - \rho_1| \, dF_\delta + \int_{1-q_1}^{1} |-1 + \rho_0 - \rho_1| \, dF_\delta \\
&= \int_0^{\rho_1} |1 + \rho_0 - \rho_1| \, dF_\delta + \int_{\rho_1}^{1-\rho_0} |\rho_0 - \rho_1| \, dF_\delta + \int_{1-\rho_0}^{1} |-1 + \rho_0 - \rho_1| \, dF_\delta \\
&= \int_0^{\rho_1} 1 + \rho_0 - \rho_1 \, dF_\delta + \int_{\rho_1}^{1-\rho_0} |\rho_0 - \rho_1| \, dF_\delta + \int_{1-\rho_0}^{1} 1 + \rho_1 - \rho_0 \, dF_\delta \, .
\end{aligned}$$

Si $\rho_0 > \rho_1$:

$$\begin{aligned}
\mathrm{TG} &= \int_0^{\rho_1} 1 + \rho_0 - \rho_1 \, dF_\delta + \int_{\rho_1}^{1-\rho_0} \rho_0 - \rho_1 \, dF_\delta + \int_{1-\rho_0}^{1} 1 + \rho_1 - \rho_0 \, dF_\delta \\
&= (1 + \rho_0 - \rho_1) \times \rho_1 + (\rho_0 - \rho_1) \times (1 - \rho_0 - \rho_1) + (1 + \rho_1 - \rho_0) \times \rho_0 \\
&= 2(\rho_0 + \rho_0\rho_1 - \rho_0^2) \, .
\end{aligned}$$

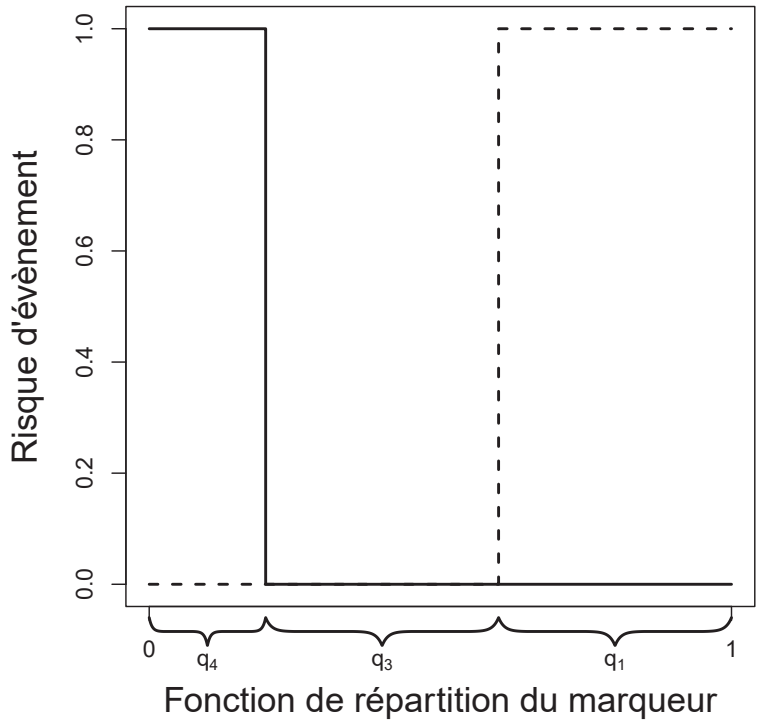


FIGURE A.1 – Exemple de courbes de risque d'un marqueur parfait pour lequel $q_2 = 0$
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

Si $\rho_0 < \rho_1$:

$$\begin{aligned} \mathrm{TG} &= \int_0^{\rho_1} 1+\rho_0-\rho_1 \, dF_\delta + \int_{\rho_1}^{1-\rho_0} \rho_1-\rho_0 \, dF_\delta + \int_{1-\rho_0}^{1} 1+\rho_1-\rho_0 \, dF_\delta \\ &= 2(\rho_1+\rho_0\rho_1-\rho_1^2)\,. \end{aligned}$$

Si $\rho_0 = \rho_1 = \rho$:

$$\mathrm{TG} = 2\rho\,.$$

En prenant maintenant le cas où $q_3 = 0$, alors $q_1 = 1-\rho_1$, $q_4 = 1-\rho_0$ et $q_2 = \rho_0+\rho_1-1$. La Figure A.2 présente un exemple de courbes de risque associées à un marqueur parfait pour lequel $q_3 = 0$.

Si $\rho_0 > \rho_1$:

$$\begin{aligned} \mathrm{TG} &= \int_0^{1-\rho_0} 1+\rho_0-\rho_1 \, dF_\delta + \int_{1-\rho_0}^{\rho_1} \rho_0-\rho_1 \, dF_\delta + \int_{\rho_1}^{1} 1+\rho_1-\rho_0 \, dF_\delta \\ &= 2(1-\rho_0+\rho_0\rho_1-\rho_1^2)\,. \end{aligned}$$

Si $\rho_0 < \rho_1$:

$$\begin{aligned} \mathrm{TG} &= \int_0^{1-\rho_0} 1+\rho_0-\rho_1 \, dF_\delta + \int_{1-\rho_0}^{\rho_1} \rho_1-\rho_0 \, dF_\delta + \int_{\rho_1}^{1} 1+\rho_1-\rho_0 \, dF_\delta \\ &= 2(1-\rho_1+\rho_0\rho_1-\rho_0^2)\,. \end{aligned}$$

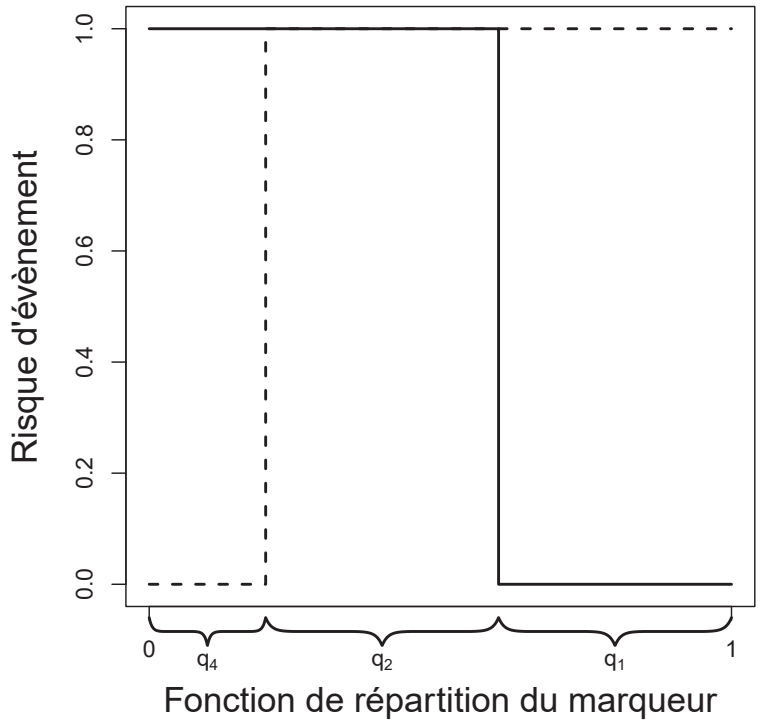


FIGURE A.2 – Exemples de courbes de risque d'un marqueur parfait pour lequel $q_3 = 0$
Courbe pleine : Traitement innovant ; Courbe en pointillés : Traitement de référence

Si $\rho_0 = \rho_1 = \rho$:

$$\mathrm{TG} = 2(1-\rho)\,.$$

Annexe B

Informations supplémentaires sur l'article relatif à l'ABC

Supplemental Material

1 Parameters of the simulation studies 1, 2 and 3

The parameters of the first, second, and third simulation studies are presented in Table S1 (for Scenario 1, the randomization constraint was satisfied in the particular case used for simulation and four Gaussian distributions could be used).

Within the context of a randomized trial, some constraints are imposed. The data should meet the randomization equation :

$$\mathrm{P}(V|T=-1) = \mathrm{P}(V|T=1)$$

where V denotes the marker. This equation can be expressed as :

$$\begin{aligned}\mathrm{P}(X_{(-1)}) \times \rho_{(-1)} + \mathrm{P}(Y_{(-1)}) \times (1-\rho_{(-1)}) \\ = \mathrm{P}(X_{(1)}) \times \rho_{(1)} + \mathrm{P}(Y_{(1)}) \times (1-\rho_{(1)})\end{aligned}$$

where $\mathrm{P}(X_{(-1)}) = P(V|T=-1, E=1)$, $\mathrm{P}(Y_{(-1)}) = P(V|T=-1, E=0)$, $\mathrm{P}(X_{(1)}) = P(V|T=1, E=1)$, $\mathrm{P}(Y_{(1)}) = P(V|T=1, E=0)$, $\rho_{(-1)} = P(E=1|T=-1)$, and $\rho_{(1)} = P(E=1|T=1)$.

In the specific case where $\mathrm{P}(X_{(-1)}) = \mathrm{P}(Y_{(1)})$, and $\mathrm{P}(X_{(1)}) = \mathrm{P}(Y_{(-1)})$, the randomization equation simplifies to:

$$\mathrm{P}(X_{(-1)})[\rho_{(-1)} - 1 + \rho_{(1)}] = \mathrm{P}(X_{(1)})[\rho_{(1)} - 1 + \rho_{(-1)}].$$

When $\rho_{(-1)} = \rho_{(1)} = 0.5$, this equation is true whatever the distributions of $X_{(-1)}$ and $X_{(1)}$. This enables the use of four Gaussian distributions in Scenario 1.

In other settings, the randomization equation enables defining $\mathrm{P}(Y_{(1)})$ as:

$$\mathrm{P}(Y_{(1)}) = \frac{\rho_{(-1)} \times \mathrm{P}(X_{(-1)}) + (1-\rho_{(-1)}) \times \mathrm{P}(Y_{(-1)}) - \rho_{(1)} \times \mathrm{P}(X_{(1)})}{1-\rho_{(1)}}$$

that does not correspond to a known theoretical distribution. For this reason, $Y_{(1)}$ values were sampled from this distribution using a Metropolis algorithm.

Table S1: Parameter settings for the first three simulation studies

Simulation study	Δ_θ	Means	Variances
First and second			
Scenario 1			
	0.1	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 0.41$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 4.57$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.41$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 4.57$
	0.2	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 4.57$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 4.57$
	0.4	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 2.21$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 2.21$
	0.6	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 1.38$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 1.38$
	0.95	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 0.71$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -1$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 0.71$
Scenario 2			
	0.1	$\mu_{X_{(-1)}} = 0.4$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 3$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.4$	
	0.2	$\mu_{X_{(-1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 3$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	
	0.4	$\mu_{X_{(-1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 1.38$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	
	0.6	$\mu_{X_{(-1)}} = 1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 0.71$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -0.64$	
Scenario 3			
	0.1	$\mu_{X_{(-1)}} = 0.32$; $\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = 0$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 1$
	0.2	$\mu_{X_{(-1)}} = 0.67$; $\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = 0$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 1$
	0.4	$\mu_{X_{(-1)}} = 0.55$; $\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = -1$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = \sigma_{Y_{(-1)}} = 1$
Third			
Scenario 1			
	0	$\mu_{X_{(-1)}} = \mu_{Y_{(1)}} = 0$	$\sigma_{X_{(-1)}} = \sigma_{X_{(1)}} = 0.71$
		$\mu_{X_{(1)}} = \mu_{Y_{(-1)}} = 0$	$\sigma_{Y_{(-1)}} = \sigma_{Y_{(1)}} = 0.71$

2 Imputation method

In the imputation method performed to deal with the censored data, the probabilities of event occurrence were estimated using the uncensored data at 21 months of follow-up with a logistic model that included the main prognostic factors of the study :

$$\text{logit}(p_i) = \beta_0 + \beta_1 \times T_i + \beta_2 \times bowelopn_i + \beta_3 \times histogrn_i + \beta_4 \times ptsn_i \times pnsn_i$$

where p_i is the probability of event occurrence in a given patient i, T_i the treatment arm, $bowelopn_i$ a binary indicator (that takes value 1 when the bowel is obstructed, otherwise 0), $histogrn_i$ the histopathological grade, $ptsn_i$ the pT pathological stage, and $pnsn_i$ the pN pathological stage.

After estimating the model parameters, the probabilities of event occurrence in patients with censored data were predicted and their outcomes sampled from a Bernoulli law.

3 Fourth simulation study - Sensitivity analysis to the equation of equal overall event risk in the two treatment arms

A fourth simulation study was performed in order to evaluate the impact of deviations from the assumption that $\rho_{(-1)} = \rho_{(1)}$ on the inference properties of the Δ_θ estimator. Two scenarios were considered, for the first one $\rho_{(-1)} = \rho_{(1)} = 0.3$ and small deviations were considered (2.5% and 5%) that led to theoretical overall risks of event equal to 0.2875 vs. 0.3125 and 0.275 vs. 0.325. For the second scenario, $\rho_{(-1)} = \rho_{(1)} = 0.1$ and the same deviations were considered which led to theoretical overall risks of event equal to 0.0875 vs. 0.1125 and 0.075 vs. 0.125.

For each of these settings, several sample sizes, N, were considered (500, 1000, 1500, and 5000). The α-risk was calculated under each setting to evaluate its inflation when the deviations from the assumption that $\rho_{(-1)} = \rho_{(1)}$ were considered. The overall mean of $\hat{\Delta}_\theta$ was also calculated and is denoted $\bar{\Delta}_\theta$.

The marker values were sampled from a Gaussian distribution so that $V \sim \mathcal{N}(0, 1)$ and the outcome of each patient was predicted using the following logistic regression model that does not include any marker-by-treatment interaction:

$$\text{logit}(p_i) = \beta_0 + \beta_V \times V_i + \beta_T \times T_i$$

The values of β_0, β_V, and β_T were chosen so as to obtain the desired overall risks of event and are presented in Table S2, and the results of the simulation study are presented in Table S3.

Table S2: Parameter settings for the fourth simulation study

$\rho_{(-1)}$ vs. $\rho_{(1)}$	β_0	β_V	β_T
0.3 vs. 0.3	-1.3846	2	0
0.2875 vs. 0.3125	-1.4805	2	0.1907
0.275 vs. 0.325	-1.579	2	0.382
0.1 vs. 0.1	-3.415	2	0
0.0875 vs. 0.1125	-3.6216	2	0.391
0.075 vs. 0.125	-3.85	2	0.791

Table S3: Results of the fourth simulation study

$\rho_{(-1)}$ vs. $\rho_{(1)}$	N	α	$\bar{\Delta}_\theta$
0.3 vs. 0.3	500	5.14%	0.00119
	1000	4.81%	0.00059
	1500	4.83%	0.00039
	5000	4.89%	0.00012
0.2875 vs. 0.3125	500	5.01%	0.00119
	1000	4.79%	0.00059
	1500	4.88%	0.00039
	5000	4.99%	0.00012
0.275 vs. 0.325	500	5.23%	0.00119
	1000	4.78%	0.00059
	1500	5.29%	0.00039
	5000	5.47%	0.00012
0.1 vs. 0.1	500	5.23%	0.00234
	1000	5.28%	0.00114
	1500	4.81%	0.00076
	5000	4.98%	0.00022
0.0875 vs. 0.1125	500	5.33%	0.00237
	1000	5.30%	0.00115
	1500	5.47%	0.00077
	5000	5.90%	0.00023
0.075 vs. 0.125	500	6.67%	0.00246
	1000	6.49%	0.00120
	1500	6.40%	0.00079
	5000	8.69%	0.00024

Annexe C

Informations supplémentaires sur l'article relatif au seuil optimal

Supplemental material

Let us consider the simple context in which the task is to decide between two treatment options, referred to as "innovative treatment" ($T = 1$) and "reference treatment" ($T = 0$). The binary event of interest is denoted by E, where $E = 1$ indicates the presence of the event of interest in a post-treatment interval, and $E = 0$ (or $\bar{E}$) indicates its absence. For example, E might be an indicator of cancer recurrence or death. Let $\rho_0 = P(E = 1|T = 0)$ and $\rho_1 = P(E = 1|T = 1)$ indicating the mean risk of event under each treatment. The candidate marker, denoted by X, is a quantitative measurement, or it may be a score that combines multiple measurements.

Details from Equation 1 to Equation 3

Under the assumptions made by Vickers et al.[1].

$$
\begin{aligned}
U(c) &= \int_c^{+\infty} f_X(x)[\rho_0(x)U_{0E}(x) + \{1-\rho_0(x)\}U_{0\bar{E}}]\,dx \\
&\quad + \int_{-\infty}^{c} f_X(x)[\rho_1(x)U_{1E}(x) + \{1-\rho_1(x)\}U_{1\bar{E}}]\,dx \\
&= U_{0E}\int_c^{+\infty} f_X(x)\rho_0(x)\,dx + U_{0\bar{E}}\int_c^{+\infty} f_X(x)[1-\rho_0(x)]\,dx \\
&\quad + U_{1E}\int_{-\infty}^{c} f_X(x)\rho_1(x)\,dx + U_{1\bar{E}}\int_{-\infty}^{c} f_X(x)[1-\rho_1(x)]\,dx \\
&= U_{0E}\rho_0[1-F_{X_{0E}}(c)] + U_{0\bar{E}}(1-\rho_0)[1-F_{X_{0\bar{E}}}(c)] + U_{1E}\rho_1 F_{X_{1E}}(c) + U_{1\bar{E}}(1-\rho_1)F_{X_{1\bar{E}}} \\
&= U_{0E}\rho_0 - U_{0E}\rho_0 F_{X_{0E}}(c) + U_{0\bar{E}}(1-\rho_0) - U_{0\bar{E}}(1-\rho_0)F_{X_{0\bar{E}}}(c) \\
&\quad + U_{1E}\rho_1 F_{X_{1E}}(c) + U_{1\bar{E}}(1-\rho_1)F_{X_{1\bar{E}}} \\
&= U_{0E}\rho_0 - U_{0E}P(E=1|X\le c, T=0)P(X\le c) + U_{0\bar{E}}(1-\rho_0) \\
&\quad - U_{0\bar{E}}P(E=0|X\le c, T=0)P(X\le c) + U_{1E}P(E|X\le c, T=1)P(X\le c) \\
&\quad + U_{1\bar{E}}P(E=0|X\le c, T=1)P(X\le c) \\
&= P(X\le c)[P(E=1|X\le c, T=0)(U_{0\bar{E}}-U_{0E}) - P(E=1|X\le c, T=1)(U_{1\bar{E}}-U_{1E}) \\
&\quad + U_{1\bar{E}} - U_{0\bar{E}}] + U_{0E}\rho_0 + U_{0\bar{E}}(1-\rho_0) \\
&= P(X\le c)[P(E=1|X\le c, T=0) - P(E=1|X\le c, T=1) + \frac{U_{1\bar{E}}-U_{0\bar{E}}}{U_{0\bar{E}}-U_{0E}}] \\
&\quad + \frac{U_{0E}\rho_0 + U_{0\bar{E}}(1-\rho_0)}{U_{0\bar{E}}-U_{0E}} \\
&= P(X\le c)[P(E=1|X\le c, T=0) - P(E=1|X\le c, T=1) - \frac{treatment}{event}] + A
\end{aligned}
$$

Details on Equation 4

As $P(X \leq c) = P(X \leq c|T = 0) = P(X \leq c|T = 1)$

$$\begin{aligned} U(c,r) &\propto F_X(c)[P(E=1|X \leq c, T=0) - P(E=1|X \leq c, T=1) - r] + A \\ &\propto F_X(c)[\frac{P(X \leq c|E=1,T=0)\rho_0}{P(X \leq c|T=0)} - \frac{P(X \leq c|E=1,T=1)\rho_1}{P(X \leq c|T=1)} - r] + A \\ &\propto P(X \leq c|E=1,T=0)\rho_0 - P(X \leq c|E=1,T=1)\rho_1 - F_X(c) \times r + A \end{aligned}$$

Then, deriving $U(c,r)$ (with respect to c) and cancelling the derivative leads to

$$\begin{aligned} U'(c,r) = 0 &\leftrightarrow f_{X_{0E}}(c)\rho_0 - f_{X_{1E}}(c)\rho_1 - f_X(c) \times r = 0 \\ &\leftrightarrow f_X(c)\rho_0(c) - f_X(c)\rho_1(c) - f_X(c) \times r = 0 \\ &\leftrightarrow \rho_0(c) - \rho_1(c) - r = 0 \\ &\leftrightarrow \rho_0(c) - \rho_1(c) = r \end{aligned}$$

Details on other scenarios and expression of the perfect marker

In the context of diagnostic or prognostic markers, in order to compare markers from different studies Baker et al.[2] proposed to standardize the utility curves by the utility associated with a perfect marker and the ones associated with extreme strategies. This standardization helps interpreting the value of the utility function on the graph, and was also proposed by Huang et al.[3] in the context of treatment selection markers. In their article Huang et al.[3] could only define the bounds of a perfect treatment selection marker. We propose to use only the upper bound of such a marker, as it is a conservative approach.

The upper bound of a perfect treatment selection marker can be derived graphically (See Figure 5) according to the decision rule, and which treatment is the more toxic. For each case, two situations exist (A or B), and depend on the mean risk of event in each treatment arm. The situation A corresponds to the case where $\rho_1 < 1 - \rho_0$, and situation B where $\rho_1 > 1 - \rho_0$.

From Figure 5, we derive the expression of the utility of the perfect treatment selection marker under each case $U_p(r)$, using the corresponding utility function.

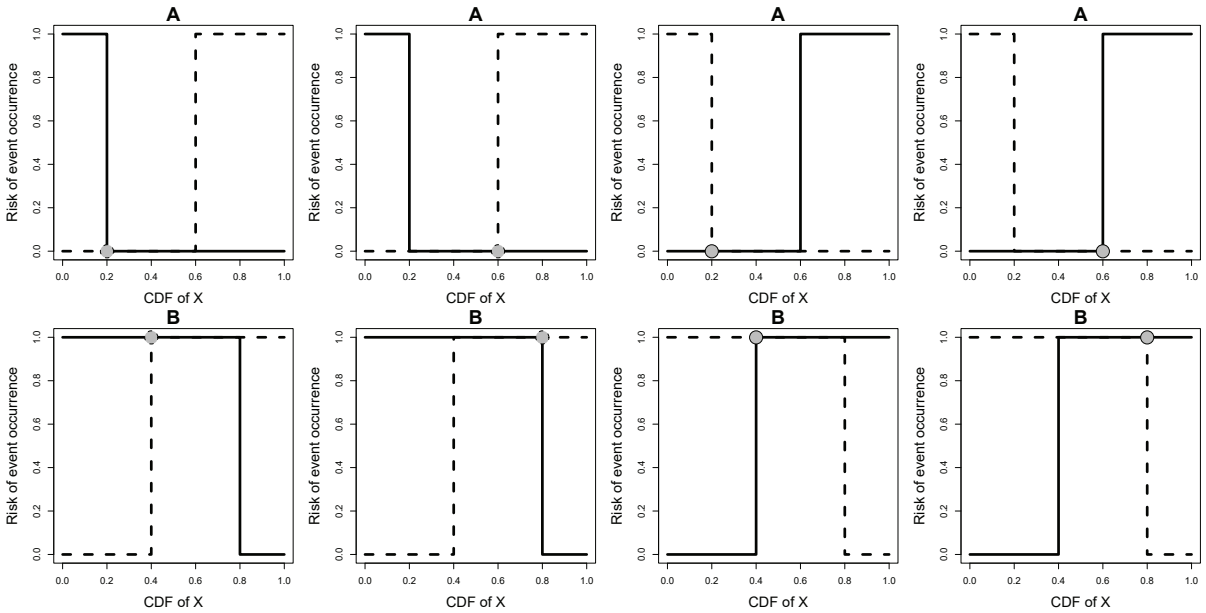


Figure 5. Risk curves associated with the perfect marker according to the decision rule, and which treatment is the most toxic. Solid line: Innovative treatment; Dotted line: Reference treatment; Grey circle: optimal threshold. First column: The reference treatment is preferred for low values of X and is the most toxic; Second column: The reference treatment is preferred for low values of X, and the innovative treatment is the most toxic; Third column: The innovative treatment is preferred for low values of X, and the reference treatment is the most toxic; Fourth column: The innovative treatment is preferred for low values of X and is the most toxic.

- The reference treatment is preferred for low values of the marker, the innovative treatment is recommended for high values of the marker, the reference treatment is the more toxic treatment option:

$$
\begin{aligned}
U(c,r) &\propto F_X(c)[P(E=1|X\leq c,T=1)-P(E=1|X\leq c,T=0)-r]+A\\
U_p(r) &= (\rho_1-[\rho_0+\rho_1-1]_+)(1-r)+A\\
U_{T=1}(r) &= \lim_{c\to+\infty} U(c,r)\propto \rho_1-\rho_0-r+A\\
U_{T=0}(r) &= \lim_{c\to-\infty} U(c,r)\propto A
\end{aligned}
$$

  where $A=-\frac{U_{1E}\rho_1+U_{1\bar{E}}(1-\rho_1)}{U_{1E}-U_{1\bar{E}}}$.
- The reference treatment is preferred for low values of the marker, the innovative treatment is recommended for high values of the marker, the innovative treatment is the more toxic treatment option:

$$
\begin{aligned}
U(c,r) &\propto F_X(c)[P(E=1|X\leq c,T=1)-P(E=1|X\leq c,T=0)+r]+A\\
U_p(r) &= (\rho_1-[\rho_0+\rho_1-1]_+)+(1-\rho_0+[\rho_0+\rho_1-1]_+)r+A\\
U_{T=1}(r) &= \lim_{c\to+\infty} U(c,r)\propto \rho_1-\rho_0+r+A\\
U_{T=0}(r) &= \lim_{c\to-\infty} U(c,r)\propto A
\end{aligned}
$$

where $A=-\frac{U_{1E}\rho_1+U_{1\bar{E}}(1-\rho_1)}{U_{1E}-U_{1\bar{E}}}$.

- The innovative treatment is preferred for low values of the marker, the reference treatment is recommended for high values of the marker, the reference treatment is the more toxic treatment option:

$$
\begin{aligned}
U(c,r) &\propto F_X(c)[P(E=1|X\leq c,T=0)-P(E=1|X\leq c,T=1)+r]+A\\
U_p(r) &= (\rho_0-[\rho_0+\rho_1-1]_+)+(1-\rho_1+[\rho_0+\rho_1-1]_+)r+A\\
U_{T=1}(r) &= \lim_{c\to+\infty} U(c,r)\propto \rho_0-\rho_1+r+A\\
U_{T=0}(r) &= \lim_{c\to-\infty} U(c,r)\propto A
\end{aligned}
$$

where $A=-\frac{U_{0E}\rho_0+U_{0\bar{E}}(1-\rho_0)}{U_{0E}-U_{0\bar{E}}}$.

- The innovative treatment is preferred for low values of the marker, the reference treatment is recommended for high values of the marker, the innovative treatment is the more toxic treatment option:

$$
\begin{aligned}
U(c,r) &\propto F_X(c)[P(E=1|X\leq c,T=0)-P(E=1|X\leq c,T=1)-r]+A\\
U_p(r) &= (\rho_0-[\rho_0+\rho_1-1]_+)(1-r)+A\\
U_{T=1}(r) &= \lim_{c\to+\infty} U(c,r)\propto \rho_0-\rho_1-r+A\\
U_{T=0}(r) &= \lim_{c\to-\infty} U(c,r)\propto A
\end{aligned}
$$

where $A=-\frac{U_{0E}\rho_0+U_{0\bar{E}}(1-\rho_0)}{U_{0E}-U_{0\bar{E}}}$,

and with $F_X(.)$ the cumulative distribution function of the marker X.

The latter situation matches with the upper bound of a perfect treatment selection marker, as defined by Huang et al.[3]

It is then possible to calculate the relative utility

$$RU(c,r) = \frac{U(c,r) - \max(U_{T=1}(r); U_{T=0}(r))}{U_p(r) - \max(U_{T=1}(r); U_{T=0}(r))}$$

The main interest of relative utility curves is to compare different markers on a common scale, and to help interpret the value of the estimated utility, 0 meaning that the marker-based strategy does not do better than one of the two extreme strategies, and 1 meaning that the marker is the perfect marker to choose between the two treatment options.

References

1. Vickers A, Kattan M and Sargent D. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials* 2007; 8(14).
2. Baker SG, Cook NR, Vickers A et al. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 2009; 172(4): 729–748.
3. Huang Y, Laber E and Janes H. Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics* 2015; 16(2): 383–399.

Annexe D

Expression analytique pour le seuil optimal et sa variance

Dans le cas où les valeurs du marqueur suivent des lois normales dans les groupes de patients $z0$ (les patients n'ayant pas développé l'évènement dans chaque bras de traitement) vec pour moyennes μ_{z0} et variances σ^2_{z0}, et supposons également que $\frac{C_Z}{C_Y} = 0$ (les traitements ont des toxicités équivalentes), alors la fonction d'utilité s'écrit sous la forme suivante :

$$U(c) \propto \Phi\left(\frac{c-\mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right) \times (1-\rho_1) - \Phi\left(\frac{c-\mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right) \times (1-\rho_0)\,,$$

avec $\Phi(.)$ la fonction de répartition d'une loi normale centrée réduite. Dans ce cas tout particulier, alors il existe une solution explicite pour le seuil optimal, c^*. Pour démontrer cela, il convient de rappeler que le seuil optimal est la valeur de marqueur qui maximise $U(c)$. Afin de maximiser la fonction $U(c)$, on annule sa dérivée :

$$\begin{aligned}
\frac{\partial U(c)}{\partial c} &= 0 \\
\frac{1}{\sqrt{2\pi}\sigma_{1\bar{Y}}} \exp\left[-\frac{1}{2}\left(\frac{c-\mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right)^2\right] \times \frac{1-\rho_1}{1-\rho_0} &= \frac{1}{\sqrt{2\pi}\sigma_{0\bar{Y}}} \exp\left[-\frac{1}{2}\left(\frac{c-\mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right)^2\right] \\
-\log\left(\sqrt{2\pi}\sigma_{1\bar{Y}}\right) - \frac{1}{2}\left(\frac{c-\mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right)^2 + \log\left(\frac{1-\rho_1}{1-\rho_0}\right) &= -\log\left(\sqrt{2\pi}\sigma_{0\bar{Y}}\right) - \frac{1}{2}\left(\frac{c-\mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right)^2 \\
\log(\sigma^2_{1\bar{Y}}) + \left(\frac{c-\mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right)^2 - \log\left(\left[\frac{1-\rho_1}{1-\rho_0}\right]^2\right) &= \log(\sigma^2_{0\bar{Y}}) + \left(\frac{c-\mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right)^2 \\
\left(\frac{c-\mu_{1\bar{Y}}}{\sigma_{1\bar{Y}}}\right)^2 - \left(\frac{c-\mu_{0\bar{Y}}}{\sigma_{0\bar{Y}}}\right)^2 + \log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right) &= 0\,.
\end{aligned}$$

Développer permet d'écrire :

$$\frac{c^2+\mu^2_{1\bar{Y}}-2c\mu_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} - \frac{c^2+\mu^2_{0\bar{Y}}-2c\mu_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} + \log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right) = 0$$

$$\frac{c^2}{\sigma^2_{1\bar{Y}}} - \frac{c^2}{\sigma^2_{0\bar{Y}}} - \frac{2c\mu_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} + \frac{2c\mu_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} + \frac{\mu^2_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} - \frac{\mu^2_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} + \log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right) = 0$$

$$\left(\frac{1}{\sigma^2_{1\bar{Y}}} - \frac{1}{\sigma^2_{0\bar{Y}}}\right)c^2 + \left(\frac{2\mu_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} - \frac{2\mu_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}}\right)c + \frac{\mu^2_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} - \frac{\mu^2_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} + \log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right) = 0\,.$$

Il s'agit ici d'un polynôme d'ordre 2, il est possible de trouver les racines de ce polynôme de la manière suivante. On pose :

$$\begin{aligned}
\Delta &= \left[2\left(\frac{\mu_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} - \frac{\mu_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}}\right)\right]^2 - 4\left(\frac{1}{\sigma^2_{1\bar{Y}}} - \frac{1}{\sigma^2_{0\bar{Y}}}\right)\left[\frac{\mu^2_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} - \frac{\mu^2_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} + \log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right)\right] \\
&= 4\left[\frac{\mu^2_{0\bar{Y}} + \mu^2_{1\bar{Y}} - 2\mu_{0\bar{Y}}\mu_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}\sigma^2_{1\bar{Y}}} + \left(\frac{1}{\sigma^2_{0\bar{Y}}} - \frac{1}{\sigma^2_{1\bar{Y}}}\right)\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right)\right] \\
&= 4\left[\frac{(\mu_{1\bar{Y}} - \mu_{0\bar{Y}})^2}{\sigma^2_{0\bar{Y}}\sigma^2_{1\bar{Y}}} + \left(\frac{\sigma^2_{1\bar{Y}} - \sigma^2_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}\sigma^2_{1\bar{Y}}}\right)\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right)\right] \\
&= \frac{4}{\sigma^2_{0\bar{Y}}\sigma^2_{1\bar{Y}}}\left[(\mu_{1\bar{Y}} - \mu_{0\bar{Y}})^2 + (\sigma^2_{1\bar{Y}} - \sigma^2_{0\bar{Y}})\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}\left[\frac{1-\rho_0}{1-\rho_1}\right]^2\right)\right].
\end{aligned}$$

Pour simplifier, on pose $R = \frac{1-\rho_0}{1-\rho_1}$, alors on écrit :

$$\Delta = \frac{4}{\sigma^2_{0\bar{Y}}\sigma^2_{1\bar{Y}}}\left[(\mu_{1\bar{Y}} - \mu_{0\bar{Y}})^2 + (\sigma^2_{1\bar{Y}} - \sigma^2_{0\bar{Y}})\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}R^2\right)\right].$$

Lorsque $\Delta \geq 0$ il existe des solutions à cette équation, autrement l'équation n'a aucune solution. Lorsque $\Delta > 0$, il existe deux solutions que l'on peut écrire sous la forme :

$$c_1 = \frac{\mu_{0\bar{Y}}\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1\right) - (\mu_{1\bar{Y}} - \mu_{0\bar{Y}}) - \frac{\sigma_{1\bar{Y}}}{\sigma_{0\bar{Y}}}\sqrt{(\mu_{1\bar{Y}} - \mu_{0\bar{Y}})^2 + \left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1\right)\sigma^2_{0\bar{Y}}\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}R^2\right)}}{\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1},$$

$$c_2 = \frac{\mu_{0\bar{Y}}\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1\right) - (\mu_{1\bar{Y}} - \mu_{0\bar{Y}}) + \frac{\sigma_{1\bar{Y}}}{\sigma_{0\bar{Y}}}\sqrt{(\mu_{1\bar{Y}} - \mu_{0\bar{Y}})^2 + \left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1\right)\sigma^2_{0\bar{Y}}\log\left(\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}}R^2\right)}}{\frac{\sigma^2_{1\bar{Y}}}{\sigma^2_{0\bar{Y}}} - 1}.$$

L'une de ces deux solutions est sous-optimale (il s'agit de la solution c_1), ainsi il est possible d'exprimer le seuil optimal c^* sous la forme :

$$c^* = \frac{\mu_{0\bar{Y}}(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\sigma^2_{0\bar{Y}}\log(b^2R^2)}}{b^2-1},$$

avec $a = (\mu_{1\bar{Y}} - \mu_{0\bar{Y}})$ et $b = \frac{\sigma_{1\bar{Y}}}{\sigma_{0\bar{Y}}}$.

Lorsque $\sigma^2_{0\bar{Y}} = \sigma^2_{1\bar{Y}} = \sigma^2_{\bar{Y}}$ alors l'expression du seuil optimal est simplifiée :

$$c^* = \frac{\sigma^2_{\bar{Y}} \log(R^2) + \mu^2_{1\bar{Y}} - \mu^2_{0\bar{Y}}}{2a}.$$

La variance du seuil optimal peut être calculée de manière analytique à l'aide de la méthode Delta :

$$\begin{aligned}\mathrm{Var}(\widehat{c}^*) \approx &\left(\frac{\partial c^*}{\partial \mu_{1\bar{Y}}}\right)^2 \mathrm{Var}(\widehat{\mu}_{1\bar{Y}}) + \left(\frac{\partial c^*}{\partial \sigma_{1\bar{Y}}}\right)^2 \mathrm{Var}(\widehat{\sigma}_{1\bar{Y}}) \left(\frac{\partial c^*}{\partial \mu_{0\bar{Y}}}\right)^2 \mathrm{Var}(\widehat{\mu}_{0\bar{Y}}) \\ &+ \left(\frac{\partial c^*}{\partial \sigma_{0\bar{Y}}}\right)^2 \mathrm{Var}(\widehat{\sigma}_{0\bar{Y}}) + \left(\frac{\partial c^*}{\partial \rho_0}\right)^2 \mathrm{Var}(\widehat{\rho}_0) + \left(\frac{\partial c^*}{\partial \rho_1}\right)^2 \mathrm{Var}(\widehat{\rho}_1).\end{aligned}$$

Pour la suite, on notera que comme les valeurs de marqueur suivent des distributions gaussiennes dans les groupes $z0$, il est possible d'écrire :

$$\widehat{\mu}_{1\bar{Y}} \sim \mathcal{N}\left(\mu_{1\bar{Y}}, \frac{\sigma^2_{1\bar{Y}}}{n_{1\bar{Y}}}\right), \qquad \widehat{\mu}_{0\bar{Y}} \sim \mathcal{N}\left(\mu_{0\bar{Y}}, \frac{\sigma^2_{0\bar{Y}}}{n_{0\bar{Y}}}\right),$$

ainsi que

$$(n_{1\bar{Y}} - 1)\frac{\widehat{\sigma}^2_{1\bar{Y}}}{\sigma^2_{1\bar{Y}}} \sim \chi^2_{n_{1\bar{Y}}-1}, \qquad (n_{0\bar{Y}} - 1)\frac{\widehat{\sigma}^2_{0\bar{Y}}}{\sigma^2_{0\bar{Y}}} \sim \chi^2_{n_{0\bar{Y}}-1}.$$

La méthode Delta permet également d'approximer les variances de $\widehat{\sigma}_{1\bar{Y}}$ et $\widehat{\sigma}_{0\bar{Y}}$, on écrit alors :

$$\begin{aligned}\mathrm{Var}(\widehat{\sigma}_{1\bar{Y}}) = \mathrm{Var}(\widehat{\sigma}^2_{1\bar{Y}})^{1/2} &\approx \left(\frac{\partial(\sigma^2_{1\bar{Y}})^{1/2}}{\partial \sigma^2_{1\bar{Y}}}\right)^2 \mathrm{Var}(\widehat{\sigma}^2_{1\bar{Y}}) \\ &= \frac{1}{4\sigma^2_{1\bar{Y}}}\mathrm{Var}(\widehat{\sigma}^2_{1\bar{Y}}) = \frac{1}{4\sigma^2_{1\bar{Y}}}\left[\frac{2(\sigma^2_{1\bar{Y}})^2}{n_{1\bar{Y}} - 1}\right] = \frac{\sigma^2_{1\bar{Y}}}{2(n_{1\bar{Y}} - 1)}.\end{aligned}$$

De la même manière on estime $\mathrm{Var}(\widehat{\sigma}_{0\bar{Y}}) \approx \frac{\sigma^2_{0\bar{Y}}}{2(n_{0\bar{Y}}-1)}$.

En posant $W = a^2 + \left(b^2 - 1\right)\sigma^2_{0\bar{Y}} \log\left(\frac{\sigma^2_{1\bar{Y}}(1-\rho_0)^2}{\sigma^2_{0\bar{Y}}(1-\rho_1)^2}\right)$, il est alors possible d'écrire explicitement $\mathrm{Var}(\widehat{c}^*)$ de la manière suivante :

$$\begin{aligned}\mathrm{Var}(\widehat{c}^*) \approx &\left\{\frac{-1 + ba \times W^{-1/2}}{b^2 - 1}\right\}^2 \times \frac{\sigma^2_{1\bar{Y}}}{n_{1\bar{Y}}} + \left\{\frac{b^2 - ba \times W^{-1/2}}{b^2 - 1}\right\}^2 \times \frac{\sigma^2_{0\bar{Y}}}{n_{0\bar{Y}}} \\ &+ \left\{\frac{2ab + (-b^2 - 1)\sqrt{W}}{(b^2-1)^2\sigma_{0\bar{Y}}} + \frac{\sigma_{1\bar{Y}} b \times \left[\log(R^2) + 1 - b^{-2}\right]}{(b^2-1)\sqrt{W}}\right\}^2 \times \frac{\sigma^2_{1\bar{Y}}}{2(n_{1\bar{Y}} - 1)} \\ &+ \left\{\frac{-2ab^2 + b(b^2+1)\sqrt{W}}{(b^2-1)^2\sigma_{0\bar{Y}}} - \frac{\sigma_{0\bar{Y}} b \times \left[\log(R^2) + b^2 - 1\right]}{(b^2-1)\sqrt{W}}\right\}^2 \times \frac{\sigma^2_{0\bar{Y}}}{2(n_{0\bar{Y}} - 1)} \\ &+ \left\{\frac{b(\sigma^2_{0\bar{Y}} - \sigma^2_{1\bar{Y}})}{(b^2-1)(1-\rho_0)\sqrt{W}}\right\}^2 \times \frac{\rho_0(1-\rho_0)}{n_0} + \left\{\frac{b(\sigma^2_{1\bar{Y}} - \sigma^2_{0\bar{Y}})}{(b^2-1)(1-\rho_1)\sqrt{W}}\right\}^2 \times \frac{\rho_1(1-\rho_1)}{n_1}.\end{aligned}$$

Résumé

En France, la recherche contre le cancer est un enjeu majeur de santé publique. On estime notamment que le nombre de nouveaux cas de cancer a plus que doublé entre 1980 et 2012. L'hétérogénéité des caractéristiques tumorales, pour un même cancer, impose des défis complexes dans la recherche de traitements efficaces. Dans ce contexte, des espoirs importants sont placés dans la recherche de biomarqueurs prédictifs reflétant les caractéristiques des patients ainsi que de leur tumeur afin d'orienter le choix de la stratégie thérapeutique. Par exemple, pour les cancers colorectaux métastatiques, il est maintenant reconnu que l'ajout de cetuximab (un anti-EGFR) à la chimiothérapie classique (ici le FOLFOX4), n'apporte un bénéfice qu'aux patients dont le gène *KRAS* est non muté. Le gène *KRAS* est ici un biomarqueur prédictif binaire, mais de nombreux biomarqueurs sont le résultat d'une quantification ou d'un dosage.
L'objectif de cette thèse est dans un premier temps, de quantifier la capacité globale d'un biomarqueur quantitatif à guider le choix du traitement. Après une revue de la littérature, une nouvelle méthode basée sur une extension des courbes ROC est proposée, et comparée aux méthodes existantes. Son principal avantage est d'être non paramétrique, et d'être indépendante de l'efficacité moyenne des traitements.
Dans un second temps, lorsqu'un biomarqueur prédictif quantitatif est étudié, la définition d'un seuil de marqueur au-delà duquel la première option de traitement sera préférée, et en-deçà duquel la deuxième option de traitement sera préférée se pose. Une approche reposant sur la définition d'une fonction d'utilité est proposée permettant alors de tenir compte de l'efficacité des traitements ainsi que de leur impact sur la qualité de vie des patients. Une méthode Bayésienne d'estimation de ce seuil optimal est proposée.

Mots-clés : évaluation de biomarqueurs ; biomarqueurs prédictifs ; courbes ROC ; seuil optimal ; fonction d'utilité ; cancer colorectal

Abstract

In France, the cancer research is a major public health issue. The number of new cancer cases nearly doubled between 1980 and 2012. The heterogeneity of the tumor characteristics, for a given cancer, presents a great challenge in the research of new effective treatments. In this context, much hope is placed in the research of predictive (or treatment selection) biomarkers that reflect the patients' characteristics in order to guide treatment choice. For example, in the metastatic colorectal cancer setting, it is admitted that the addition of cetuximab (an anti-EGFR) to classical chemotherapy (the FOLFOX4), only improve the outcome of patients with *KRAS* wild-type tumors. In that context, the *KRAS* gene is a binary treatment selection marker, but plenty of biomarkers result from some quantifications or dosage measurements.
The first aim of this thesis is to quantify the global treatment selection ability of a biomarker. After a review of the existing litterature, a method based on an extension of ROC curves is proposed and compared to existing methods. Its main advantage is that it is non-parametric, and that it does not depend on the mean risk of event in each treatment arm.
In a second time, when a quantitative treatment selection biomarker is assessed, there is a need to estimate a marker thereshold value above which one treatment is preferred, and below which the other treatment is recommended. An approach that relies on the definition of a utility function is proposed in order to take into account both efficacy and toxicity of treatments when estimating the optimal threshold. A Bayesian method for the estimation of the optimal threshold is proposed.

Keywords: biomarker evaluation ; predictive biomarker ; treatment selection ; receiver operating characteristic curves ; optimal threshold ; utility function ; colorectal cancer

Laboratoire de Biométrie et Biologie Evolutive - Equipe Biostatistiques-Santé
162 avenue Lacassagne
69424 Lyon Cedex 03