Méthodologie de l'utilisation des biomarqueurs quantitatifs longitudinaux pour l'aide à la décision en médecine. Application aux PSA dans le cancer de la prostate

Fabien Subtil

N° d'ordre : 72-2010 Année 2010

Thèse

présentée devant l'Université Claude Bernard Lyon 1

pour l'obtention du

Diplôme de Doctorat

(*arrêté du 7 août 2006*)

Soutenue publiquement le 4 juin 2010

par

Fabien Subtil

Méthodologie de l'utilisation des biomarqueurs quantitatifs longitudinaux pour l'aide à la décision en médecine
Application aux PSA dans le cancer de la prostate

Composition du jury

Directeur de thèse	Professeur René ECOCHARD, Université Lyon 1
Co-encadrant	Docteur Muriel RABILLOUD, Université Lyon 1
Rapporteurs	Professeur Louis Rachid SALMI, Université Bordeaux 2
	Monsieur Martyn PLUMMER, CIRC
Examinateur	Docteur Pascal GIRARD, Université Lyon 1

UNIVERSITE CLAUDE BERNARD – LYON 1

Président de l'Université	**M. le Professeur L. Collet**
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine Lyon Sud - Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. Le Professeur F. Gieres
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. Lieto
Institut Universitaire de Technologie A	Directeur : M. le Professeur C. Coulet
Institut Universitaire de Technologie B	Directeur : M. le Professeur R. Lamartine
Institut de Science Financière et d'Assurance	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M R. Bernard

Remerciements

Mes remerciements vont tout d'abord à René Ecochard et Muriel Rabilloud, qui m'ont accueilli dans leur équipe, initié à la recherche, et encadré tout au long de cette thèse. Merci de m'avoir proposé un sujet, parfois difficile, mais tellement enrichissant.

Mes remerciements vont également aux rapporteurs de cette thèse, Messieurs Louis-Rachid Salmi et Martyn Plummer, pour leurs remarques très constructives, ainsi qu'à Monsieur Pascal Girard, pour avoir accepté de faire partie du jury.

Je remercie la Ligue Nationale contre le Cancer dont le soutien financier m'a permis d'effectuer cette thèse dans de bonnes conditions.

Merci aux enseignants de la filière Bioinformatique et Modélisation de l'INSA, sans qui je n'aurais peut-être jamais fait cette thèse.

Un grand merci à tous les membres du Service de Biostatistique pour leurs conseils et leur amitié, et plus particulièrement à ceux situés à proximité de mon bureau. Merci pour votre soutien, votre enthousiasme et votre bonne humeur. Merci Alvine, Pierre, Sylvain, Jean, Arnaud, Jean-Damien, Stéphanie, Mathieu et Michèle.

Enfin, merci infiniment à ma famille, mes parents et ma sœur, pour votre soutien constant, et votre affection inestimable.

Table des matières

Table des figures

Liste des tableaux

Introduction

La thématique générale de ce travail de recherche concerne la méthodologie d'utilisation des biomarqueurs mesurés de manière répétée au cours du suivi de patients, pour le diagnostic précoce de maladie ou le pronostic de typologies d'évolution et pour l'aide à la décision médicale. Le suivi des patients s'effectue, pour certains types de maladies, par le dosage régulier de certaines substances dans l'organisme, ainsi que par des examens cliniques. La question est de savoir comment utiliser au mieux ces successions de mesures pour prendre une décision concernant une éventuelle adaptation de la stratégie thérapeutique (Slate et Turnbull, 2000 ; Proust-Lima et Taylor, 2009). L'objectif est d'identifier les personnes pouvant bénéficier d'un traitement, traitement considéré ici au sens large.

Les mesures répétées au cours du temps permettent la construction d'un profil d'évolution du biomarqueur, profil pouvant être décrit par le niveau de ce marqueur aux différents temps, mais également par d'autres critères, tels que la vitesse de progression (pente), ou la survenue de modifications de cette vitesse de progression (rupture de pente). Dans un premier temps, il est nécessaire de choisir le critère issu du biomarqueur qui est le plus adapté pour cibler les personnes pouvant bénéficier du traitement. Ce critère étant bien souvent quantitatif, il faut définir ensuite un seuil au dessus ou en dessous duquel le traitement est à recommander. La détermination de ce seuil et le choix du critère ne se fondent pas uniquement sur la capacité à discriminer les patients, mais également sur le bénéfice et le coût, en termes d'état de santé, à diagnostiquer ou prédire à tort ou à raison la survenue d'un événement (Jund et *al.*, 2005). A chacune de ces étapes sont associés un certain nombre de problèmes méthodologiques, auxquels des réponses sont apportées dans ce travail.

Les méthodes développées sont appliquées à des données concernant le suivi de patients traités par ultrasons en raison d'un cancer de la prostate, afin de détecter la persistance de cellules cancéreuses après le traitement (Blana et *al.*, 2009). Très souvent, une biopsie est réalisée de façon systématique dans les trois à six mois suivant le traitement pour diagnostiquer la persistance de cancer. Une option consisterait à utiliser les dosages répétés d'antigènes spécifiques de la

prostate (PSA) réalisés au cours du suivi afin de ne déclencher une biopsie que lorsqu'il y a de fortes chances qu'elle soit positive. L'objectif est donc de déterminer comment utiliser au mieux les mesures de PSA pour ne déclencher une biopsie que lorsqu'il y a de fortes chances qu'elle soit positive.

La première partie de ce document décrit les problématiques à l'origine du travail de recherche, à la fois cliniques (chapitre un) et méthodologiques (chapitre deux).

La seconde partie du manuscrit traite de la façon de choisir un critère reflétant la cinétique du biomarqueur afin de discriminer les patients. Le chapitre trois revient sur les différentes méthodes d'évaluation des critères pour le diagnostic ou le pronostic. Le chapitre quatre présente l'article publié dans Statistics in Medicine, dans lequel est proposée une méthode pour estimer de façon robuste ces critères à partir des mesures répétées de PSA.

La troisième partie de la thèse revient sur l'estimation du seuil optimal du critère à partir d'une méthode Bayésienne décrite dans un article accepté dans le Biometrical Journal (chapitre cinq). Le sixième chapitre insiste sur l'importance de tenir compte des préférences des patients et des cliniciens en termes d'état de santé dans la détermination du seuil du critère ; cette réflexion a fait l'objet d'un article rédigé pour The European Urology.

Enfin, la dernière partie du manuscrit présente un certain nombre de perspectives.

Bien que ce document corresponde à une thèse d'articles, de très nombreux compléments sont apportés à ceux-ci. Les principes généraux des méthodes qu'ils décrivent sont souvent représentés dans ce document, afin d'insister sur des détails particuliers, ou de leur donner un autre éclairage.

Première partie

Contexte et problématiques

Chapitre

1

Problématiques de santé

1.1 L'ère des biomarqueurs

1.1.1 Qu'est-ce qu'un biomarqueur ?

Bien que le terme biomarqueur soit relativement récent, les biomarqueurs sont utilisés depuis de nombreuses années dans les études pré-cliniques et cliniques. Ils peuvent être définis comme des caractéristiques mesurables objectivement et évaluées comme des indicateurs d'un processus biologique normal, d'un processus pathogénique, ou bien d'une réponse pharmacologique à une intervention thérapeutique (Biomarkers Definitions Working Group., 2001). La plupart des biomarqueurs sont des substances biologiques dont la quantité est mesurée dans certaines parties de l'organisme, comme le sang ou des tissus spécifiques. Les produits issus des gènes, les anticorps, les enzymes ou encore les hormones en sont des exemples. De façon plus générale, le degré d'activité d'une fonction organique complexe ou les caractéristiques d'une structure biologique peuvent également être considérés comme des biomarqueurs.

L'aide au diagnostic de maladies a grandement bénéficié de la découverte de biomarqueurs. Depuis un certain nombre d'années, le terme biomarqueur est très fortement associé aux nouvelles méthodes de recherche moléculaire, portant sur l'expression des protéines, des gènes, ou encore sur l'analyse du polymorphisme nucléotidique. Ransohoff (2004) parle d'ailleurs de la " grande promesse " des nouveaux marqueurs pour le diagnostic de certaines maladies.

Il existe désormais des biomarqueurs dans de très nombreuses spécialités, tout particulièrement en oncologie (Schiffer, 2009) et dans le domaine des maladies cardiovasculaires

(Gerszten et Wang, 2008). Des exemples classiques sont le dosage des antigènes CA19-9 ou CA125 pour le diagnostic du cancer du pancréas, ou bien le suivi des taux de lymphocytes T CD4 dans le sang pour diagnostiquer le passage au stade SIDA (syndrome d'immunodéficience acquise). Tous les jours, des publications annoncent la découverte de nouveaux biomarqueurs. Néanmoins, il convient de rester prudent, car beaucoup d'entre eux se révèlent inutiles lorsqu'ils sont utilisés à l'échelle de la population (Ransohoff, 2004). Ainsi, une attention importante doit être accordée au développement des biomarqueurs, ainsi qu'à leur évaluation.

1.1.2 A quoi sert un biomarqueur en médecine ?

Les utilisations des biomarqueurs en médecine sont très nombreuses. Ils peuvent permettre de détecter une maladie dans sa phase précoce de développement dans une population asymptomatique ; on parle alors de biomarqueur utile au *dépistage*. Pour des patients qui présentent déjà des signes ou symptômes d'une maladie, certains biomarqueurs permettent d'établir de manière définitive la présence de la maladie ; ce sont des biomarqueurs utiles au *diagnostic*. Très souvent, ils sont utilisés afin d'éviter un examen clinique invasif. Lorsque des symptômes de la maladie sont présents, mais que celle-ci n'est pas encore détectable cliniquement au moment de la mesure du biomarqueur, on parle alors de *diagnostic précoce*. D'autres biomarqueurs sont utilisés lorsqu'une maladie a déjà été diagnostiquée, comme par exemple un cancer, et que l'on souhaite en prédire la typologie d'évolution : ce sont des biomarqueurs utiles pour le *pronostic*. Face à une maladie, différents traitements sont parfois envisageables et le choix peut être effectué en fonction d'un biomarqueur permettant de prédire la réponse probable aux traitements ; on parle de biomarqueur utile pour le *choix de traitement*. Enfin, l'efficacité d'un traitement peut être mesurée grâce à un biomarqueur utile pour *la surveillance active*. Dans ce travail de thèse, on s'intéressera aux biomarqueurs pour le dépistage, le diagnostic ou le diagnostic précoce et pour le pronostic. Pour ne pas alourdir la présentation, les termes " malades " et " non malades " seront utilisés de manière générique, même si dans certains cas, il faudrait parler de futurs malades ou de typologie d'évolution. De même, il sera considéré de manière générique que l'objectif de l'utilisation des biomarqueurs est de prendre une décision de traitement, même si la décision ne correspond pas toujours à un traitement.

Dans certains cas, la présence d'un biomarqueur dans l'organisme, quelle que soit sa concentration, est révélatrice de la maladie ; par exemple, un produit synthétisé spécifiquement par un virus ne peut être détecté chez une personne que si le virus est présent ; quelle que soit la quantité de produit mesurée, elle signe la présence du virus de manière spécifique, correspondant

ainsi à un signe pathognomonique. On parle alors de biomarqueur binaire ou *qualitatif*. Dans d'autres cas, c'est une concentration élevée (ou faible) du biomarqueur qui est le signe de la maladie. On parle de biomarqueur *quantitatif*. Néanmoins, l'utilisation d'un biomarqueur quantitatif comme un test diagnostique ou pronostique nécessite la définition d'un seuil au dessus (ou en dessous) duquel la concentration est jugée anormale et évocatrice de la maladie ; avec ce seuil de positivité, le résultat associé à la mesure du biomarqueur redevient binaire. Un signe pathognomonique peut également être quantitatif. Par exemple, la détection dans le plasma d'ARN spécifique du virus d'immunodéficience humaine (VIH) atteste de la présence de la maladie ; mais la concentration d'ARN, appelée charge virale, est également informative. En effet, cette dernière est étroitement corrélée à l'évolution de la maladie et constitue un outil prédictif fiable du risque de développer un SIDA.

Des procédures diagnostiques plus élaborées peuvent combiner le résultat de plusieurs biomarqueurs et les informations cliniques concernant le patient. Cette combinaison peut s'utiliser de deux façons. La première repose sur un algorithme décisionnel, consistant à ordonner les différentes informations, puis à les analyser successivement, l'interprétation donnée au résultat d'une information dépendant de l'interprétation de l'information précédente. La seconde façon d'utiliser une combinaison de biomarqueurs ou d'informations cliniques consiste à convertir ces différentes informations en un score global sur une échelle quantitative ; il faut alors à nouveau définir un seuil au dessus ou en dessous duquel la valeur du score est considérée comme évocatrice de la présence de la maladie. Par la suite, le terme *marqueur* sera employé de manière générale, un marqueur pouvant être un biomarqueur, ou le résultat d'une combinaison de biomarqueurs ou de caractéristiques de l'individu.

A une valeur donnée du marqueur quantitatif peut être associé un risque de maladie. Pour les tests diagnostiques, un risque de 80 % indique que le patient a une probabilité de 80 % d'avoir la maladie. Pour les tests de dépistage ou pour le diagnostic précoce, un risque de 80 % indique que la probabilité que l'individu développe la maladie est de 80 %. Dans certains cas, une notion de durée peut être introduite ; le risque est alors calculé pour différents niveaux du marqueur et différentes durées. Ainsi, pour une valeur donnée du marqueur et une durée de deux ans, un risque de 80 % indique que l'individu a 80 % de chances de développer la maladie durant les deux années suivant la mesure du marqueur. Pour le pronostic, un risque de 80 % indique que la maladie du patient a 80 % de chances d'évoluer dans un type d'état, dans un délai fixé.

1.1.3 Les biomarqueurs dynamiques

Pour certaines maladies, une seule mesure du biomarqueur permet d'effectuer un diagnostic ; on parlera de biomarqueur *fixe*. Dans d'autre cas, notamment pour le diagnostic précoce ou le dépistage, le biomarqueur est mesuré de manière répétée au cours du temps et c'est l'évolution de ces valeurs qui peut être le reflet du développement de la maladie ; on parlera alors de biomarqueur *dynamique*. Pour Zolg et Langen (2004), " un biomarqueur est une molécule qui indique une altération de l'état physiologique d'un individu en relation avec son état de santé ou de maladie. D'après cette définition, un biomarqueur n'est pas statique, mais il change au cours du temps ".

Deux types d'études sur les biomarqueurs dynamiques sont à distinguer en fonction de la façon dont le statut du patient est défini. Certaines études tiennent compte du fait que le statut du patient change au cours du temps ; ce dernier n'est pas malade au début du suivi, mais le deviendra éventuellement. Dans ce cas, il est intéressant d'étudier l'évolution de la capacité du marqueur à prédire la détection clinique de ce changement de statut en fonction du délai entre la mesure du biomarqueur et la détection effective. Intuitivement, le biomarqueur devrait prédire d'autant mieux la détection clinique de la maladie que la mesure en est proche dans le temps. La question est alors de savoir jusqu'à combien de jours, mois ou années avant la mise en évidence clinique du changement de statut la capacité du marqueur à détecter ce changement est encore acceptable. Cette question est cruciale dans le cas du dépistage, ou lorsqu'il faut un certain temps avant que les thérapies soient efficaces.

D'autres études ne tiennent pas compte de l'évolution du statut du patient au cours du temps et ne considèrent que son état final. Le patient est soit " destiné " à développer la maladie, soit ne la développera jamais. Ainsi, le statut du patient n'est pas le statut au moment de la mesure, mais un statut latent, représentant son évolution future ; ce type de définition se retrouve dans les modèles dits de guérison (Berkson et Gage, 1952). Par exemple, pour des patientes ayant subi une aspiration en raison d'une maladie trophoblastique, une des questions est de savoir si, en cas de persistance de mole après aspiration, celle-ci risque de devenir cancéreuse. L'objectif est alors d'amorcer un traitement le plus tôt possible. L'évolution du taux d'hormone hCG après l'aspiration peut permettre de répondre à la question précédente. Dans ce cas, les cliniciens souhaitent classer les patientes en deux groupes : celles totalement guéries et celles pour lesquelles une mole cancéreuse sera détectée, sans forcément distinguer les patientes pour lesquelles la détection sera précoce et celles pour lesquelles elle sera tardive. Un autre exemple

en oncologie est le dosage répété des antigènes CA-125 pour prédire la réponse à un traitement en raison d'un cancer des ovaires (Rustin, 2003). Un exemple similaire dans le domaine de l'immunologie est la détection de la résistance à des traitements (Nevirapine, Delavirdine et Efavirenz) pour le VIH, détection basée sur la mesure régulière de la charge virale (Kohlmann et *al.*, 2009). Dans ces exemples, l'objectif est de créer deux groupes : le groupe dit des malades et celui des non malades, sans forcément distinguer dans le groupe des malades les échecs au traitement précoces ou tardifs. Bien souvent, ces études en sont à des stades préliminaires. D'autres études plus approfondies sont réalisées, par la suite, pour analyser la capacité du biomarqueur à classer les patients en fonction du délai entre la mesure et la détection clinique de la maladie. Dans ce travail de thèse, on s'intéressera plus particulièrement au premier type d'études, mais les méthodes développées pourraient être adaptées au second.

Lorsque le biomarqueur est dynamique, il faut définir la façon d'utiliser la succession de mesures effectuées chez un même patient pour prédire au mieux son statut futur ou latent. Le niveau du marqueur à un temps donné peut à lui seul suffire, mais ce n'est pas toujours le cas. Pour le cancer des ovaires, le taux d'antigène CA-125 à un temps donné suivant le traitement est peu différent entre les patientes en échec de traitement et celles pour lesquelles le traitement est efficace. Ceci est dû, entre autres, au fait que les taux de CA-125 au moment du traitement sont très variables. Dans ce cas, le pourcentage de diminution est plus informatif que le taux lui-même. Pour la maladie trophoblastique, le niveau des hCG à un temps donné suivant l'aspiration chez les femmes avec persistance de la maladie est également très peu différent de celui chez les femmes guéries, mais certaines caractéristiques reflétant la cinétique des hCG varient d'un groupe à l'autre. La décroissance des hCG après aspiration s'effectue en deux étapes : une première étape de décroissance très rapide, suivie d'une étape de diminution plus lente. Plusieurs " marqueurs " de la persistance de la maladie trophoblastique persistante sont envisageables, comme la vitesse de décroissance lors de la première ou de la deuxième phase. Ici, le marqueur n'est plus le biomarqueur en lui-même, mais une des caractéristiques reflétant la cinétique du biomarqueur.

Des méthodes sont nécessaires afin de comparer les différents marqueurs reflétant la cinétique d'un biomarqueur dynamique et choisir, ensuite, celui le plus approprié pour distinguer les patients selon leur statut latent. Ceci sera l'objet des chapitres trois et quatre. Ces méthodes ont été appliquées au diagnostic de persistance du cancer de la prostate après un premier traitement par ultrasons.

1.2 Les PSA et le cancer de la prostate

1.2.1 Le cancer de la prostate et ses traitements

De la grosseur d'une noix, la prostate est une glande du système reproducteur masculin située sous la vessie, en avant du rectum. Elle sécrète les substances nutritives et fluidifiantes du sperme. Cette glande peut être le site de diverses affections : l'infection de la prostate, l'inflammation de la prostate (ou prostatite), l'adénome de la prostate ou hypertrophie bénigne et enfin le cancer de la prostate. Avec plus de 40 000 cas détectés par an en France, le cancer de la prostate est la seconde cause de mortalité par cancer chez l'homme. Il représente plus de 10 000 décès par an (Bauvin et *al.*, 2003) et constitue au delà de 70 ans la première cause de décès par cancer. Environ un homme sur neuf présentera une forme clinique de cette maladie dans sa vie. Le cancer de la prostate apparaît rarement avant 40-50 ans et la plupart des cas sont constatés entre 60 et 90 ans, avec un âge médian de détection de 74 ans (Grosclaude et *al.*, 2007).

Beaucoup de cancers de la prostate sont maintenant diagnostiqués à un stade précoce et sont localisés ; un traitement curatif est donc possible. Néanmoins, le cancer de la prostate pouvant avoir une évolution lente, une simple surveillance peut parfois suffire. Dans le cas d'un traitement, le choix de ce dernier est fonction du stade du cancer, de sa vitesse de progression et des caractéristiques du patient. Le traitement par chirurgie radicale, ou prostatectomie, est le traitement de référence en cas de cancer localisé. Il permet un contrôle de la maladie cancéreuse dans environ 75 % des cas (Hull et *al.*, 2002), mais il n'est pas dépourvu de morbidité (Lowrance et *al.*, 2010). Le traitement du cancer de la prostate par curiethérapie (ou brachythérapie) est également recommandé pour les cancers localisés. Il consiste à irradier la prostate grâce à des grains radioactifs insérés sous contrôle échographique. L'efficacité est plus faible que celle de la prostatectomie, mais les effets secondaires sont moins nombreux (Ciezki et Klein, 2009). Le traitement par radiothérapie externe est quant à lui envisageable pour des cancers localisés ou à l'état de métastase. Il est réservé aux cancers de stade élevé, en raison d'une toxicité non négligeable (Ciezki et Klein, 2009). Pour les cancers hormono-sensibles, les patients dont le cancer n'est plus localisé à la prostate peuvent également être traités par hormonothérapie ; l'objectif n'est alors plus de guérir le cancer, mais de diminuer la progression de la maladie. L'hormonothérapie peut également être indiquée en cas d'échec d'un traitement local, comme l'ablation de la prostate ou la radiothérapie, ou en association à ces mêmes traitements.

1.2.2 Le traitement UFHI

Un traitement alternatif est le traitement par ultrasons focalisés de haute intensité (UFHI). Il repose sur l'émission d'ultrasons par une sonde endoréctale focalisée sur la prostate. Tout comme les rayons solaires passant au travers d'une loupe peuvent développer une grande chaleur au point de focalisation, les ultrasons passant au travers de différents tissus produisent au point focal une chaleur intense (entre 80 et 100°C). La destruction tissulaire dans la zone cible est provoquée par la combinaison de trois phénomènes :

- la nécrose de coagulation, correspondant à la destruction instantanée et définitive des cellules en raison de l'absorption de l'énergie ultrasonore par le tissu ;
- la cavitation, qui est liée à la mise en vibration de microbulles de gaz très chaudes par les impulsions ultrasonores successives ; ces microbulles entraînent également la destruction des tissus ;
- la diffusion de chaleur, liée non pas aux tirs d'ultrasons, mais à leur répétition toutes les quatre secondes ; la chaleur ainsi engendrée se propage sur plusieurs millimètres autour du volume cible proprement dit.

Le traitement UFHI est efficace pour les patients ayant un cancer de la prostate localisé, de risque faible ou intermédiaire, avec des effets indésirables ou complications thérapeutiques modérés (Blana et *al.*, 2008a). L'Association Française d'Urologie le reconnaît désormais comme une option de traitement pour les patients de plus de 70 ans, ayant un cancer de stade T1 ou T2, avec un taux de PSA inférieur à 15 ng/mL et un score de Gleason inférieur à 7 (Blana et *al.*, 2008b). Ce traitement est donc particulièrement envisageable pour les personnes âgées, pour lesquelles l'état de santé ne permet pas une opération chirurgicale, ou encore pour les personnes qui ne souhaitent pas de prostatectomie. Des taux de survie sans récidive à 5 ans de 66 à 77 % ont été mentionnés dans la littérature (Blana et *al.*, 2009 ; Poissonnier et *al.*, 2007a ; Uchida et *al.*, 2009). Poissonnier et *al.* (2007a) rapportent un taux d'incontinence tardive de 9 %, un taux de sténose urétrale de 6 % et un taux d'impuissance de 31 à 61 %, en fonction de la préservation ou non des bandelettes neurovasculaires.

Un des avantages majeurs du traitement UFHI est qu'il peut être répété en cas d'échec (Poissonnier et *al.*, 2007a), contrairement au traitement par radiothérapie qui abîme trop le tissu prostatique. Il peut également être utilisé comme traitement de sauvetage après une radiothérapie (Murat et *al.*, 2009), ou à l'inverse, une radiothérapie peut être réalisée comme traitement de sauvetage après un traitement UFHI (Pasticier et *al.*, 2008). Les effets secondaires

et complications thérapeutiques augmentent tout de même avec le nombre de traitements (Poissonnier et *al.*, 2007a). Un second traitement par ultrasons est d'autant plus efficace qu'il est effectué de manière précoce (Murat et *al.*, 2010), d'où la nécessité de marqueurs permettant de détecter rapidement la persistance du cancer après le premier traitement.

1.2.3 La détection du cancer de la prostate

La biopsie prostatique, avec observation au microscope des cellules prostatiques, est l'examen permettant de poser de manière définitive le diagnostic de cancer de la prostate. Une biopsie négative ne permet pas de rejeter le diagnostic de cancer, un patient pouvant avoir des cellules cancéreuses, mais pas dans la zone où le prélèvement a été effectué. Néanmoins, des techniques récentes de biopsies guidées par imagerie par résonance magnétique permettent de détecter avec beaucoup plus de certitude un cancer (Rouvière et *al.*, 2010). L'inconvénient majeur de la biopsie est lié à son caractère invasif. Si le pourcentage de complications est modéré (9 % d'après Lee et *al.* (2009)), le nombre élevé d'examens entraîne un nombre non négligeable de complications. Des examens non invasifs sont donc souhaitables pour effectuer des biopsies uniquement lorsque les suspicions de cancer sont élevées.

Le toucher rectal est l'un des outils du médecin afin de surveiller l'apparition d'un cancer de la prostate, et permettant de déclencher, dans le cas d'un toucher rectal anormal, des examens plus poussés. Depuis les années 1980, le dosage de l'antigène spécifique de la prostate (PSA) est un second outil à la disposition du médecin. Les PSA sont des protéines synthétisées par la glande prostatique ; un de leurs rôles reconnus est de fluidifier le liquide prostatique. En temps normal, leur concentration dans le sang est faible. Une désorganisation du tissu prostatique, retrouvée dans le cas du cancer de la prostate, est à l'origine d'un passage plus important de PSA dans la circulation générale. Néanmoins, leur augmentation marque simplement une anomalie de la prostate, qui peut être due à une inflammation, à un adénome de la prostate, ou à une intervention chirurgicale et non nécessairement à un cancer de la prostate. Tous ces facteurs, ainsi que la variabilité de précision de mesures entre les laboratoires d'analyses et la variabilité biologique naturellement importante des PSA, peuvent entraîner des variations d'une mesure à l'autre de l'ordre de 30 à 40 % sans aucun rapport avec la présence éventuelle d'un cancer (Eastham et *al.*, 2003 ; Soletormos et *al.*, 2005).

L'Association Française d'Urologie recommande un dosage annuel de PSA associé à un toucher rectal pour les hommes entre 50 et 75 ans. Un taux supérieur à 4 ng/mL est reconnu comme anormal, même si cette valeur commence à être critiquée dans la littérature (Cao et

Yao, 2010). L'utilisation des PSA pour le dépistage systématique du cancer de la prostate est de plus en plus polémique. Deux essais randomisés de taille conséquente ont été réalisés sur la mortalité due au cancer de la prostate, avec dépistage systématique par PSA ou non, l'un en Europe (Schröder et *al.*, 2009), l'autre aux Etats-Unis (Andriole et *al.*, 2009). Les résultats intermédiaires de l'étude européenne montrent un léger effet du dépistage systématique en faveur d'une réduction de la mortalité, ceux de l'étude américaine ne montrent aucun effet. Et même si un léger effet existe, son utilité en termes d'état de santé reste à prouver, car de très nombreuses biopsies négatives sont réalisées en raison du dépistage systématique (Schröder et *al.*, 2009). Ainsi, il est de moins en moins reconnu que le dosage des PSA constitue un outil pour le dépistage du cancer de la prostate.

L'utilisation des PSA pour la détection d'une récidive locale du cancer de la prostate porte beaucoup moins à controverse. Suite à un traitement, tel que la radiothérapie, la curiethérapie ou le traitement UFHI, une forte diminution des PSA est observée, suivie par une phase de stabilisation ou de ré-augmentation. Après radiothérapie, le critère Phoenix est maintenant couramment reconnu pour définir un échec biochimique du traitement (Roach et *al.*, 2006). Un patient est considéré comme en échec en cas :

- d'une valeur de PSA mesurée au dessus du nadir plus 2 ng/mL, le nadir étant défini comme la plus basse valeur de PSA atteinte au cours du suivi du patient ;
- d'une biopsie positive ;
- de l'administration d'un traitement de sauvetage.

D'autres critères reposant également sur les PSA après traitement ont été proposés, même s'ils n'ont pas été retenus par la communauté scientifique :

- un temps de doublement du taux de PSA réduit (Daskivich et *al.*, 2006) ;
- trois augmentations successives du taux de PSA après le nadir (ASTRO, 1997) ;
- une valeur de nadir de PSA supérieure à 0,2 ng/mL (Critz et *al.*, 1999 ; Ray et *al.*, 2006a), ou à 1,5 ng/mL (Zelefsky et *al.*, 2009).

Dans le cas du traitement par ultrasons, il n'existe pas encore de consensus sur le critère à retenir pour définir un échec de traitement. L'objectif de l'application de ce travail de thèse a été d'établir un critère, issu des mesures successives de PSA, qui permette de discriminer les patients en termes de persistance de cellules cancéreuses ou non, limitant ainsi le nombre de biopsies réalisées. Ces critères ne sont pas des critères pronostiques, mais des critères permettant de prendre la décision de réaliser une biopsie. Même s'il sera dit par la suite, de manière générique, que les biomarqueurs permettent de prendre une décision de traitement, dans le cadre des

données de PSA après traitement UFHI, l'objectif n'est pas directement de prendre la décision de retraiter, mais d'abord la décision de réaliser une biopsie.

1.2.4 La cohorte de patients du service d'urologie et de chirurgie de la transplantation de l'Hôpital Edouard Herriot

Les équipes de l'INSERM et du service d'urologie et de chirurgie de la transplantation de l'Hôpital Edouard Herriot à Lyon ont effectué de nombreuses recherches pour la mise au point du traitement UFHI pour le cancer de la prostate. Depuis, le service d'urologie dispose d'une des plus grandes cohortes de patients traités par UFHI. C'est au sein de cette cohorte que l'étude sur les critères permettant de déterminer la persistance du cancer de la prostate après le traitement a été menée, en retenant les patients traités depuis janvier 2000 et en arrêtant le suivi à janvier 2008. Cette étude avait pour but d'estimer et de comparer les performances diagnostiques de différents critères issus des mesures longitudinales de PSA, puis d'estimer le seuil optimal du meilleur critère.

1.2.4.1 Suivi des patients

Après le traitement UFHI, tous les patients ont été suivis avec des dosages de PSA dans le sang, en moyenne toutes les trois semaines durant les quatre premiers mois, puis tous les mois jusqu'au huitième et ensuite tous les quatre mois. Les concentrations de PSA sont exprimées en ng/mL. La plupart des patients ont eu une biopsie sextante de contrôle entre le troisième et le sixième mois et éventuellement des biopsies supplémentaires selon la pratique courante des cliniciens du service d'urologie. Les biopsies ont été réalisées à chaque fois par l'un des quatre radiologistes expérimentés, selon une procédure standardisée définie au sein de l'unité ; les radiologistes ne disposaient pas de l'information concernant les mesures de PSA au moment de l'interprétation des résultats des biopsies.

1.2.4.2 Sélection des patients

L'étude a été restreinte aux patients répondant à quatre critères :

– le traitement UFHI était le premier traitement effectué pour le cancer de la prostate, excluant ainsi les patients ayant déjà eu un traitement UFHI, une radiothérapie, une cryothérapie, ou une chirurgie de la prostate (pour ces patients, le traitement UFHI correspond à un traitement de sauvetage) ;

– aucune hormonothérapie n'avait été administrée avant le traitement UFHI ;
– la durée de suivi du patient était au minimum de 90 jours ;
– le patient avait eu une indication de biopsie après 90 jours de suivi, selon la pratique courante des cliniciens du service d'urologie.

L'évaluation des performances diagnostiques de critères issus des mesures longitudinales de PSA pour détecter la persistance locale de cellules cancéreuses après le traitement UFHI nécessitait un examen indépendant des PSA, permettant d'attester de manière certaine de la présence ou de l'absence de cellules cancéreuses. L'examen le plus couramment utilisé est la réalisation d'une biopsie sextante de la prostate, une biopsie positive étant forcément synonyme de la persistance de cellules cancéreuses, à l'exception des biopsies effectuées durant les trois premiers mois. En effet, ces dernières peuvent détecter des cellules cancéreuses détruites par le traitement ultrasons, mais pas encore éliminées par l'organisme. Dans le cas de la radiothérapie, il est conseillé de ne pas interpréter les résultats de biopsie avant douze à dix-huit mois (Crook et *al.*, 1995) ; pour le traitement UFHI, l'élimination des cellules cancéreuses est plus rapide, et les biopsies sont couramment jugées interprétables dès trois mois (Rouvière et *al.*, 2010). Ainsi, les résultats des biopsies effectuées durant les 90 premiers jours n'ont pas été pris en compte dans l'analyse. Si pour les premiers patients de la cohorte, une biopsie était réalisée systématiquement entre trois et six mois, cette pratique a évolué au cours du temps, les biopsies n'étant par la suite déclenchées qu'en cas de suspicion d'échec de traitement. Ces suspicions d'échec étaient en partie fondées, pour certains patients, sur l'évolution des valeurs de PSA. Ainsi, un certain nombre de patients ont été retirés de l'analyse car ils ne disposaient pas de biopsie après trois mois de suivi.

Le fait de ne pas disposer de biopsie au delà de trois mois de suivi est en faveur de l'absence de cellules cancéreuses résiduelles. Mais un patient peut avoir une biopsie positive alors que ses valeurs de PSA sont stables et peu élevées. Ainsi, les patients n'ayant pas eu de biopsie au delà de trois mois de suivi ne pouvaient pas être considérés de manière certaine comme des patients non en échec de traitement ; les inclure comme des " non malades " dans l'analyse aurait pu entraîner une sous estimation de la spécificité des critères diagnostiques issus des PSA. Ainsi, au risque de fournir des estimations biaisées de performances de critères a été préféré celui de limiter la portée des résultats de l'étude uniquement aux patients pour lesquels une indication de biopsie après 90 jours serait effectuée d'après la pratique des cliniciens du service d'urologie de l'hôpital Edouard Herriot entre 2000 et 2008. Cette population correspond à une population plus à risque de persistance de cellules cancéreuses.

Six patients ont été supprimés de l'analyse car le protocole de traitement UFHI ne correspondait pas au protocole de traitement standard défini au sein de l'unité.

Pour des raisons calculatoires, les patients ne disposant pas d'au moins cinq mesures de PSA et d'une mesure de PSA après 90 jours de suivi ont été retirés de l'analyse.

La fréquence des mesures de PSA dépend en partie de l'évolution de ces mesures, une valeur anormalement élevée lors d'un examen entraînant en règle générale un nouvel examen plus rapproché. Ainsi, les patients n'ayant pas eu plus de quatre mesures de PSA au cours de leur suivi étaient vraisemblablement des patients moins à risque d'être en échec de traitement.

Les patients avec persistance de cancer après le traitement – les malades – ont été définis comme étant les patients ayant eu au moins une biopsie positive, les patients restants ayant été considérés comme les non malades. Le fait que toutes les biopsies soient négatives ne certifie pas l'absence de cellules cancéreuses, mais le résultat des biopsies était l'unique source d'information disponible pour définir le statut des patients. Ainsi, l'étude de la capacité des PSA à détecter la persistance du cancer de la prostate a été remplacée par l'étude de la capacité des PSA à discriminer les patients ayant au moins une biopsie positive de ceux n'en ayant pas, la combinaison des biopsies étant une mesure imparfaite du vrai statut latent du patient. L'étude et l'ensemble des résultats ont été restreints aux patients pour lesquels les cliniciens ont une suspicion de persistance locale de cellules cancéreuses après trois mois de suivi.

La population étudiée a ainsi été réduite à 289 patients, dont 139 n'ont eu aucune biopsie positive (les non malades) et 150 au moins une biopsie positive (les malades). La figure 1.1 représente le processus de sélection des patients de l'étude.

1.2.4.3 Description de la cohorte

Les caractéristiques des patients de l'étude au moment du traitement UFHI sont décrites dans le tableau 1.1. Parmi ces patients, 7 ont été perdus de vue et 15 sont décédés. Les patients ont eu, en moyenne, 8,2 mesures de PSA, les quartiles des nombres de mesures étant de 6, 7 et 9. 69,5 % des patients n'ont eu qu'une biopsie (202 patients), 26 % deux biopsies (74 patients), 4 % trois biopsies (12 patients) et 0,5 % quatre biopsies (1 patient). La plupart des cancers étaient peu ou moyennement aggresifs, les stades cliniques étant majoritairement T1 et T2 et les score de Gleason inférieurs ou égal à 7 pour la plupart des patients. Le temps de suivi moyen, en nombre de jours écoulé depuis le traitement UFHI, était de 876 jours (écart type de 730 jours) ; les quartiles de ce temps de suivi étaient de 263, 622 et 1321 jours. La moyenne des dates de

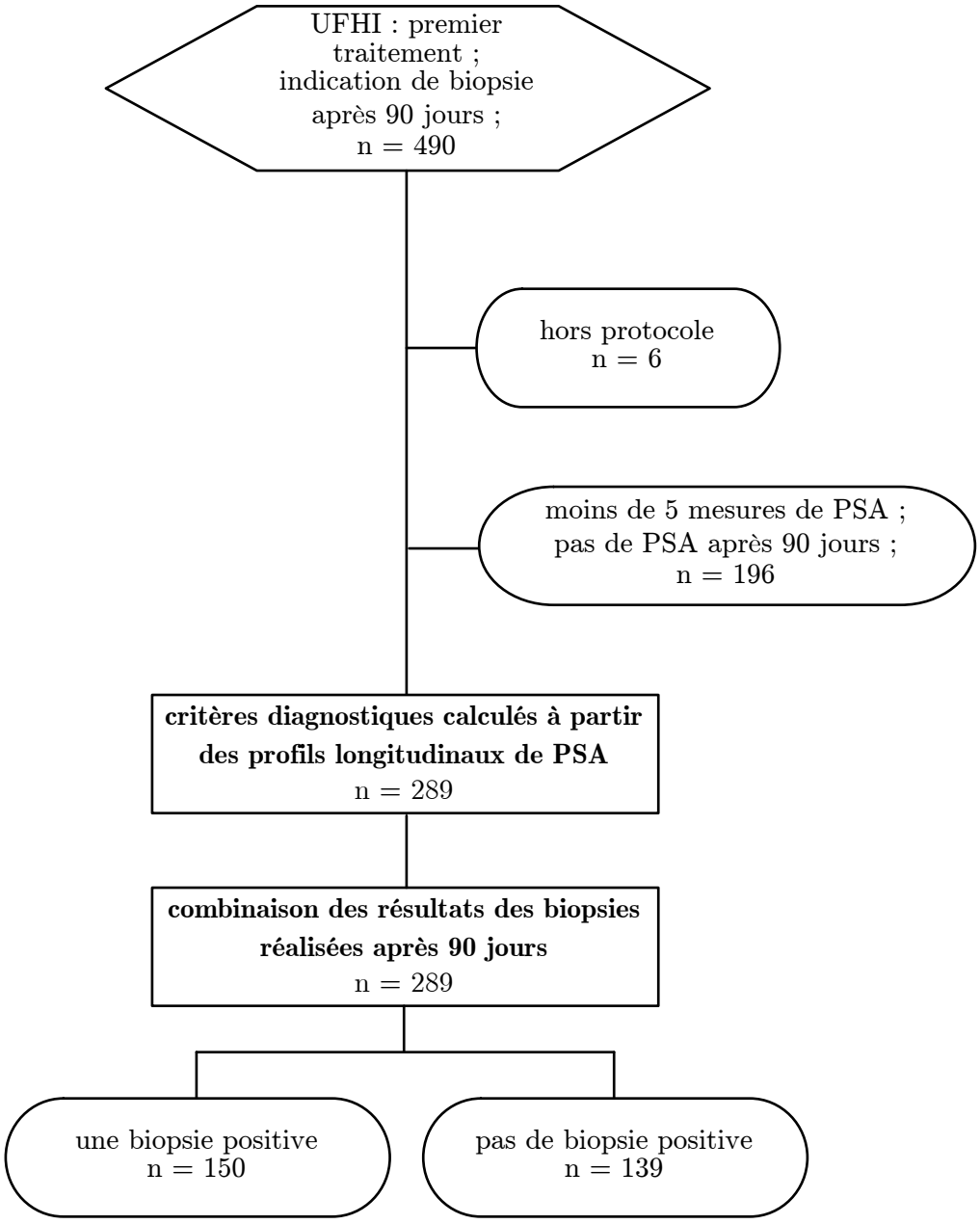

Figure 1.1 – Définition des patients de l'étude.

Caractéristique	
Age moyen (médian) au traitement	69,7 ± 5,6 (71) années
Taux moyen (médian) de PSA pré-traitement	8,25 ± 6,91 (7,15) ng/mL
Stade clinique du cancer	
T1	156 patients (54 %)
T2	124 patients (43 %)
T3	9 patients (3 %)
Score de Gleason du cancer	
non défini	5 patients (2 %)
≤ 6	164 patients (57 %)
$= 7$	95 patients (33 %)
≥ 8	25 patients (8 %)
Volume prostatique moyen (médian)	27,73 ± 13,62 (27) mL

± : écart type

Tableau 1.1 – Caractéristiques des patients de l'étude.

biopsie postive était de 504,9 jours (écart type de 492,8 jours) ; les quartiles de ces dates étaient de 151, 269 et 753 jours.

1.2.5 PSA et statut du patient

Les valeurs de PSA en fonction de la date de la mesure par rapport au traitement UFHI sont représentées sur les figures 1.2 et 1.3, avec à gauche les profils des patients non malades et à droite ceux des malades. Ces profils sont caractérisés par une très forte diminution des PSA juste après le traitement, puis par une phase de stabilisation ou de ré-augmentation.

Durant la phase de stabilisation ou de ré-augmentation, les valeurs de PSA des malades n'étaient globalement pas beaucoup plus élevées que celles des non malades. Le niveau de PSA à une date donnée n'est donc pas le meilleur critère pour discriminer les deux groupes, c'est pourquoi l'étude a porté sur des critères reflétant la cinétique des PSA. Trois critères mentionnés dans la littérature ont été analysés :

- le nadir, plus faible valeur de PSA atteinte ; plus il est élevé, plus le risque de biopsie positive est élevé ;
- la date du nadir (Ray et *al.*, 2006b) ; plus le nadir est atteint tardivement, plus le risque de biopsie positive est élevé ;

– la vélocité des PSA, ou vitesse de ré-augmentation après la phase de diminution (Blana et *al.*, 2009) ; plus elle est élevée, plus le risque de biopsie positive est élevé.

Le premier objectif de la thèse a été de comparer la capacité des trois marqueurs (nadir, date du nadir et vélocité) à discriminer les patients en termes de résultats de biopsies attendus, pour des patients suspectés d'être en échec de traitement après trois mois de suivi (chapitre 4). Une fois le marqueur choisi, il a fallu choisir un seuil au delà duquel le patient est jugé malade et au delà duquel une biopsie est fortement conseillée (chapitre 5). Ces questions, bien que simples en apparence, posent de nombreux problèmes méthodologiques, décrits par la suite.

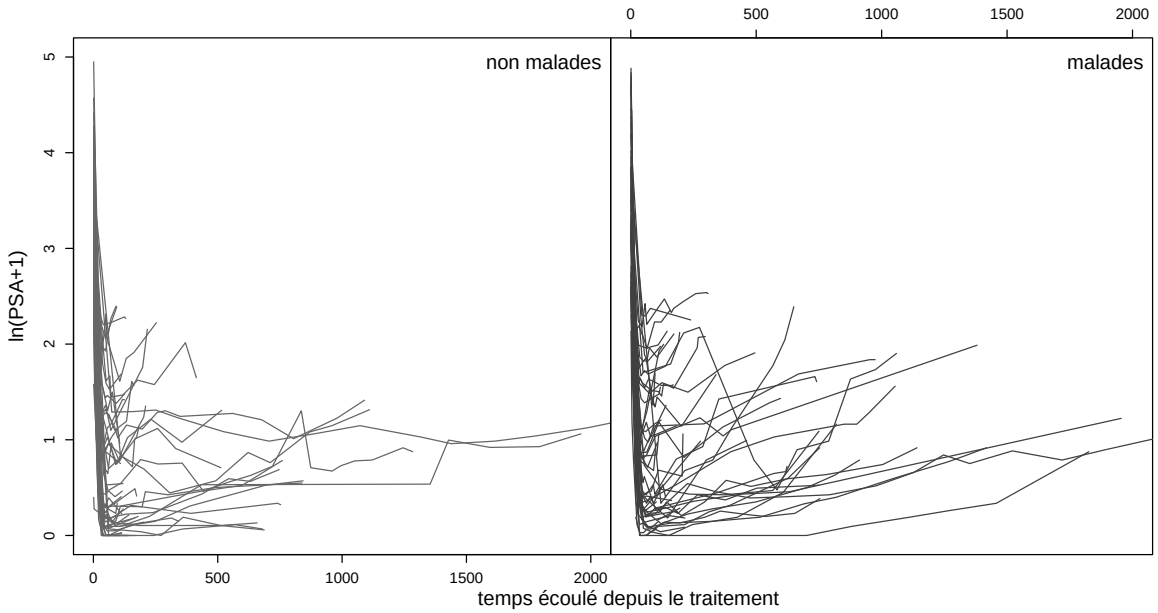

Figure 1.2 – Valeur de ln(PSA+1) (ln(ng/mL)) en fonction du temps écoulé depuis le traitement (jours).

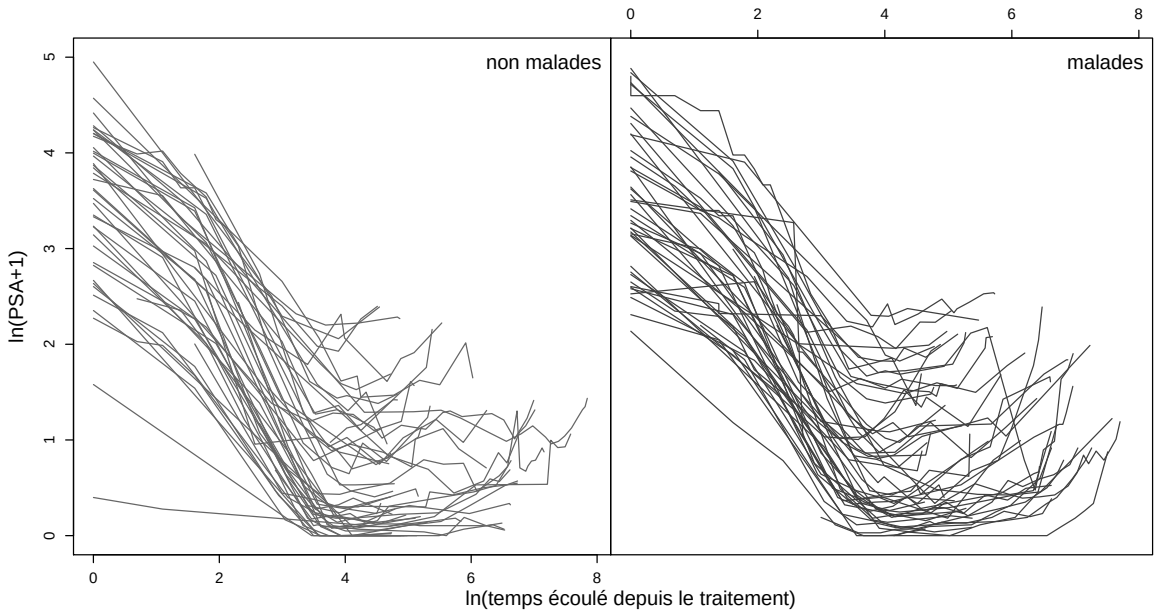

Figure 1.3 – Valeur de ln(PSA + 1) en fonction du logarithme du temps écoulé depuis le traitement (ln(jours)).

Chapitre

2 Problématiques méthodologiques

Lorsqu'un biomarqueur dynamique est utilisé pour distinguer deux groupes de patients et prendre une décision vis à vis d'eux, une première étape consiste à choisir le marqueur, reflétant la cinétique du biomarqueur, qui est le plus approprié pour cette tâche. Pour cela, il est nécessaire d'estimer les différents marqueurs pour chacun des patients, puis de comparer leur capacité à discriminer. Ces marqueurs prenant très souvent des valeurs continues, il faut donc ensuite définir un seuil du meilleur marqueur au dessus ou en dessous duquel le patient est jugé malade, ou allant développer la maladie.

2.1 Estimation des marqueurs

2.1.1 Marqueur empirique, marqueur modélisé

Dans le cadre des données de PSA, il est relativement facile de calculer de façon empirique le nadir, la date du nadir et la vélocité. Le nadir est la plus basse valeur de PSA atteinte et la vélocité est la pente des PSA après le nadir. C'est d'ailleurs de cette façon que le nadir est calculé habituellement. Néanmoins, les taux de PSA ne sont pas mesurés tous les jours, alors que leurs variations sont très rapides durant les premières semaines suivant le traitement UFHI (figure 1.2). Il faut donc distinguer le *nadir théorique*, qui serait la plus basse valeur de PSA réellement atteinte si des mesures étaient effectuées par exemple toutes les heures et s'il n'y avait pas d'erreur de mesure, et le *nadir mesurable* ou *nadir clinique*, qui est la plus basse valeur de PSA effectivement mesurée. Par exemple, le cas d'un patient dont la vraie valeur de nadir est

de 0,1 ng/mL est considéré. La dernière mesure de PSA avant ce nadir est de 0,5 ng/mL, car la décroissance des PSA n'est pas terminée à cette date et la mesure suivante est de 0,3 ng/mL, la ré-augmentation des PSA ayant déjà commencé ; pour ce patient, le nadir théorique est donc de 0,1 ng/mL et le nadir mesurable de 0,3 ng/mL. La fréquence des mesures a, par conséquent, un impact sur le nadir mesurable.

Les erreurs de mesures et la variabilité biologique naturelle des PSA peuvent également entraîner des variations importantes entre le nadir théorique et le nadir mesurable. Le patient peut, au moment où il est censé atteindre son nadir, subir une infection, ou présenter une inflammation, entraînant une remontée temporaire des PSA. Cette augmentation n'est pas en lien avec la persistance de cancer et le nadir ainsi mesuré est plus élevé que celui que le patient aurait atteint s'il n'avait pas eu d'infection. Le nadir théorique sera considéré par la suite comme la plus basse valeur de PSA qu'aurait pu atteindre un patient, en l'absence d'erreur de mesure et de phénomènes autres que la persistance de cancer.

La date du nadir mesuré est également liée à la fréquence des mesures et aux erreurs de mesure. De même, la vélocité dépend fortement des erreurs de mesure ou de phénomènes intercurrents autres, tels une ré-augmentation des PSA. Une valeur anormalement élevée peut beaucoup influencer la pente estimée après le nadir. L'importance de l'écart entre les valeurs mesurées et les valeurs théoriques sera quantifiée dans le cadre des données de PSA.

Dans une première phase d'étude où l'objectif est de comparer plusieurs marqueurs issus d'un biomarqueur dynamique, il est important d'utiliser des valeurs de marqueurs qui soient indépendantes de la fréquence des mesures et des erreurs de mesure, ou de phénomènes qui ne dépendent pas du processus étudié, ici la persistance du cancer. En effet les conclusions quant au choix du marqueur risquent de dépendre de ces phénomènes intercurrents et ne pas refléter uniquement la capacité des marqueurs à discriminer les patients. Ainsi, le choix du marqueur doit reposer sur les valeurs théoriques des marqueurs par patient.

Une méthode pour s'affranchir, en partie, du problème de la fréquence des mesures consiste à modéliser l'évolution du biomarqueur dynamique – on parlera par la suite du *profil* d'évolution du biomarqueur – puis à calculer les valeurs de marqueurs à partir des valeurs modélisées du biomarqueur. Il sera montré, dans le cadre des données de PSA, que les marqueurs modélisés sont moins biaisés que les marqueurs mesurés. De plus, en utilisant des méthodes de modélisation robuste, il est possible de limiter l'impact des valeurs aberrantes dues à des phénomènes intercurrents, tels une infection temporaire.

Cette méthode de modélisation est utile uniquement durant la phase d'évaluation et de choix du marqueur, ainsi que pour l'estimation de la valeur limite du marqueur retenu pour déclencher une action. Dans la pratique courante, les marqueurs sont calculés directement à partir des valeurs de biomarqueurs mesurés, et surtout, les cliniciens ne disposent pas de tout le suivi du patient pour prendre une décision. Ainsi, il est important de comparer les performances diagnostiques obtenues à partir des valeurs de marqueurs modélisées à celles obtenues avec les valeurs mesurées de marqueur, afin de s'assurer que les méthodes développées conduisent à des résultats acceptable si elles sont utilisées en pratique courante, avec les valeurs mesurées de marqueurs.

2.1.2 Modélisation des profils de biomarqueur

2.1.2.1 Choix du modèle

La modélisation des profils de biomarqueur nécessite dans un premier temps le choix d'un modèle adapté. Dans le cadre de mesures répétées au cours du temps, le modèle correspond à un modèle dit longitudinal. Le choix de la forme du modèle pour décrire un type de profil fait appel à la fois à des arguments provenant de la connaissance biologique du phénomène et à des arguments empiriques d'adéquation entre le modèle et les données.

La qualité des estimations de valeurs de marqueurs dépend en grande partie de l'adéquation entre le modèle et les données. Ainsi, il faut être vigilant à ne pas choisir un modèle imposant des hypothèses trop contraignantes. Par exemple, l'écart entre la valeur observée et la valeur modélisée est souvent supposé suivre une distribution normale, ce qui est rarement le cas dans la réalité, particulièrement pour les données de PSA, en raison des augmentations de PSA temporaires inexpliquées. Les mesures associées à des augmentations brusques de PSA, non liées à la persistance de cellules cancéreuses, seront par la suite appelées des *valeurs aberrantes*. Il a été montré que pour les modèles non-linéaires – les biomarqueurs dynamiques ayant souvent une évolution non-linéaire avec le temps – l'inférence sur des paramètres basée sur une loi normale pour décrire les écarts entre l'observé et le prédit est vulnérable à la présence de valeurs aberrantes (Lange et *al.*, 1989 ; Pinheiro et *al.*, 2001). Une méthode est décrite dans le chapitre quatre pour tenir compte de ces valeurs aberrantes dans l'estimation des paramètres du modèle.

2.1.2.2 Modèle à effets mixtes

Une fois le modèle choisi, il faut ajuster celui-ci sur l'ensemble des patients. Une première solution consiste à ajuster le modèle séparément pour chacun d'eux. Ceci nécessite entre autres de disposer de suffisamment de mesures par patient pour que l'estimation des paramètres soit possible et que l'ajustement fourni soit correct. Une autre solution, particulièrement utile lorsque certains patients ont peu de mesures, consiste à utiliser un modèle à effets mixtes (Laird et Ware, 1982).

Supposons que l'on dispose de n patients, y_{ij} correspondant à la $j^{\text{ième}}$ mesure des m_i mesures de biomarqueur du patient i, mesure effectuée au temps t_{ij}. L'espérance de la valeur de PSA pour une date de mesure et un patient donné est supposée être donnée par une fonction ψ, caractérisant le profil d'évolution du biomarqueur et faisant intervenir les dates des mesures et des covariables. Parmi les paramètres du modèle, certains peuvent avoir un effet identique quel que soit le patient ; pour d'autres, il peut exister des variations des valeurs d'effet entre les patients. Par exemple, un paramètre décrivant la vitesse de ré-augmentation des PSA après le nadir varie d'un patient à l'autre autour d'une moyenne, chaque patient ayant sa propre vitesse de ré-augmentation. Le modèle à effets mixtes fait donc intervenir plusieurs types de paramètres : les paramètres à effet fixe dont l'effet ne varie pas d'un patient à l'autre, notés $\boldsymbol{\alpha}$, et les paramètres à effet aléatoire, notés $\mathbf{b_i}$, donc l'effet varie d'un patient à l'autre, d'où la mention de l'indice i. Ainsi, pour un patient i, le vecteur des mesures $\mathbf{y_i}$ peut être décrit par la relation suivante :

$$\mathbf{y_i} = \psi(\mathbf{X_i}, \boldsymbol{\alpha}, \mathbf{Z_i}, \mathbf{b_i}) + \boldsymbol{\varepsilon_i}$$

$\mathbf{X_i}$ correspond aux covariables associées aux paramètres à effet fixe, $\mathbf{Z_i}$ à celles associées aux paramètres à effet aléatoire et $\boldsymbol{\varepsilon_i}$ correspond au vecteur des résidus (écart entre l'observé et le prédit). La variabilité des effets aléatoires d'un patient à l'autre est caractérisée par une loi ; dans la plupart des cas, les effets aléatoires sont supposés suivre une loi normale multivariée sur l'ensemble des patients :

$$\mathbf{b_i} \hookrightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

où $\boldsymbol{\Sigma}$ est une matrice de variance covariance. Lorsque la moyenne d'un effet aléatoire sur l'ensemble des patients est non nulle, alors cet effet moyen est introduit dans les paramètres à effet fixe. Les effets aléatoires sont supposés être indépendants des résidus.

Enfin, en imposant que les effets aléatoires suivent une distribution sur l'ensemble des patients, les paramètres d'un individu sont estimés à la fois à partir de ses propres mesures, mais également à partir des mesures des autres patients, en donnant d'autant moins de poids aux données des autres patients que la quantité d'informations contenue dans les mesures du patient en question est élevée. Les modèles à effets mixtes sont donc particulièrement adaptés lorsque certains patients ne disposent pas de suffisamment de mesures pour estimer leurs paramètres.

2.1.2.3 Modélisation robuste

Pour les modèles à effets mixtes, les effets aléatoires sont couramment supposés suivre des distributions gaussiennes. Ceci simplifie entre autres les calculs. L'inférence concernant les effets fixes est robuste quant au choix de la distribution des effets aléatoires (Verbeke et Lesaffre, 1997 ; Chen et *al.*, 2002), avec une perte d'efficacité si la distribution choisie n'est pas adaptée. Par contre, l'inférence sur les effets aléatoires n'est pas robuste à une mauvaise spécification de leur distribution (Kleinman et Ibrahim, 1998), or, pour de nombreux modèles longitudinaux, les effets aléatoires ne suivent pas tous des lois normales. Dans le cadre des données de PSA, la qualité des valeurs de marqueurs estimées par patient dépend de la qualité des estimations des effets aléatoires, mais il existe d'autres situations où les estimations des effets aléatoires sont également intéressantes en elles-mêmes (Tsiatis et *al.*, 1995 ; Ohlssen et *al.*, 2007).

Ainsi, il est nécessaire de développer des méthodes robustes, permettant d'assouplir à la fois l'hypothèse de non-normalité des effets aléatoires et l'hypothèse de non-normalité des résidus dans le cadre des modèles à effets mixtes. Ceci est décrit dans le chapitre quatre.

2.2 Comparaison de marqueurs

Une fois les marqueurs estimés de la façon la plus fiable possible pour chaque patient, il faut comparer leur capacité à détecter correctement le statut latent du patient, afin de choisir le meilleur marqueur. Une mesure de la capacité à détecter le statut des patients est nécessaire ; l'expression " mesure de performance " sera employée par la suite, en la laissant volontairement floue.

2.2.1 Mesure des performances de marqueurs

L'objectif est, à partir d'un échantillon de patients dont le statut vis à vis de la maladie est connu, de déterminer la capacité du marqueur à classer de nouveaux patients en futurs malades

et futurs non malades et d'analyser la généralisabilité du marqueur à d'autres patients. Ceci correspond à la *classification supervisée*, par opposition à la *classification non supervisée*, pour laquelle le statut des patients observés n'est pas connu, son objectif étant de mettre en évidence des groupes de patients aux caractéristiques communes (Hand, 1997). Une hypothèse implicite de la classification supervisée est que la distribution dont seront issus les nouveaux patients est similaire à celle de l'échantillon de patients étudié. Cette hypothèse n'est pas toujours vérifiée, entraînant dans ce cas une dérive de la population.

Déterminer le meilleur marqueur nécessite une mesure de performance du marqueur, mais celle-ci dépend de ce qui est entendu par meilleur marqueur. Le terme " meilleur " indique qu'un critère à définir est à optimiser. Il peut y avoir plusieurs raisons d'utiliser un marqueur et ces différentes raisons se traduisent vraisemblablement par différentes mesures de performance (Hand, 1997). De très nombreuses mesures ont été décrites dans la littérature. L'objectif, ici, n'est pas de toutes les exposer, mais de les regrouper par catégories. La présentation de ces mesures met en évidence des oppositions fortes ; dans la réalité, ces mesures ne sont pas opposées, mais combinées, le choix d'un marqueur n'étant jamais déterminé uniquement par une seule de ces mesures. De plus, il est à noter que les outils statistiques associés à ces différentes échelles de mesures ne sont volontairement pas présentés en même temps que les échelles, la définition des échelles ne dépendant pas des outils ; de plus, certains outils ne sont pas forcément rattachés à une seule échelle. La présentation simultanée des échelles et des outils risquerait d'être confuse. Ainsi, plusieurs échelles sont décrites ci-après, la présentation des outils associés étant reportée au chapitre trois.

2.2.2 Le paradigme de la calibration

Pour un malade, ou son médecin, l'objectif est d'avoir une information sur l'état de santé présent ou futur, afin de pouvoir prendre une décision. D'après la valeur du marqueur, il est possible de calculer la probabilité que le patient ait la maladie, ou le risque qu'il la développe. La probabilité ou le risque sont compris entre zéro et un ; cette échelle est la même quel que soit le marqueur, ce qui facilite l'interprétation des résultats obtenus. En fonction de la probabilité de maladie, de l'état de santé de l'individu et de son aversion vis-à-vis de la maladie ou des traitements, le clinicien et le patient prennent une décision. Il n'y a pas uniquement deux actions possibles, comme traiter ou ne pas traiter, mais un ensemble d'actions envisageables : choix parmi plusieurs traitements, mise sous surveillance active, examens complémentaires... Lorsque le marqueur est en cours de développement, toutes les actions possibles suite à sa mesure ne sont

pas forcément identifiées ; les décisions sont prises au cas par cas, en fonction des caractéristiques du patient.

L'évaluation du marqueur est effectuée ici indépendamment des conséquences des décisions qui seront envisagées suite à la mesure, les décisions étant prises au cas par cas. L'objectif est uniquement d'apporter une information la plus juste sur l'état de santé actuel ou futur du patient d'après la valeur du marqueur, sans se soucier de l'utilisation qui sera faite de cette information. Ainsi, une mesure de performance est l'adéquation entre le risque prédit pour un niveau du marqueur et le vrai risque du patient. Cette mesure est appelée la *calibration*.

Il est important de noter que, bien que le risque soit présenté ici comme apportant une information à l'échelle individuelle, le modèle permettant d'associer une valeur du marqueur à un risque est construit sur un échantillon de la population.

L'information apportée par le biomarqueur sur le risque réside dans l'écart des probabilités de valeurs de marqueur entre les groupes malades et non malades. Soit Y la valeur du marqueur et M le vrai statut du patient vis à vis de la maladie ($M = 0$ pour un patient non malade et $M = 1$ pour un patient malade ou qui va le devenir). A l'aide du théorème de Bayes, la probabilité que l'individu soit malade, d'après la valeur du marqueur ($P(M = 1|Y)$), est donnée par :

$$P(M = 1|Y) = \frac{P(Y|M = 1) \times P(M = 1)}{P(Y)}$$

$P(Y|M = 1)$ correspond à la probabilité de la valeur de marqueur Y dans le groupe des malades. Le rapport entre la probabilité que l'individu soit malade et celle qu'il ne le soit pas, sachant la valeur du marqueur, est obtenue à l'aide de la relation suivante :

$$\frac{P(M = 1|Y)}{P(M = 0|Y)} = \frac{P(Y|M = 1)}{P(Y|M = 0)} \times \frac{P(M = 1)}{P(M = 0)} \tag{2.1}$$

Dans cette formule, $P(M = 1)/P(M = 0)$ correspond à l'information disponible a priori sur le statut du patient, sans connaissance de la valeur du biomarqueur ; cette information est obtenue à partir de la prévalence de la maladie dans la population dont est issu le patient. Cette connaissance est modifiée, grâce au marqueur, au travers du ratio $P(Y|M = 1)/P(Y|M = 0)$, appelé ratio de vraisemblance. Plus les écarts de probabilité de valeurs de marqueur entre les deux groupes sont importants, plus ce ratio prend des valeurs extrêmes (élevées ou faible selon les valeurs du marqueur), plus le biomarqueur apporte par conséquent d'information sur le risque du patient, en plus de l'information a priori issue de la prévalence.

Pour Moons et Harrell (2003), l'évaluation d'un marqueur passe par la modélisation du risque de maladie prédit, car ce qui importe au patient est son risque en fonction du niveau du biomarqueur.

2.2.3 Le paradigme de la discrimination

Lorsqu'une nouvelle stratégie de traitement ou de prise en charge d'une maladie est mise au point, l'objectif n'est pas le traitement d'un unique patient, mais d'un ensemble de patients présentant des caractéristiques similaires ; il faut dont définir quels sont les patients, parmi la population étudiée, qui pourront retirer un bénéfice du traitement. Les patients en question sont ceux qui présentent ou qui présenteront réellement la maladie. Un marqueur est nécessaire pour les distinguer de ceux qui ne la présenteront pas, afin de n'administrer le traitement qu'aux personnes qui en sont la cible. Les décideurs, qui peuvent être par exemple un ensemble de médecins, choisissent donc un marqueur permettant de prendre une décision pour un ensemble de patients. Il est souhaitable que ce marqueur soit bénéfique pour le plus grand nombre possible de personnes. L'objectif, ici, est donc de créer deux groupes de patients, la formation de ces groupes dépendant des décisions qui seront prises en fonction de l'appartenance à l'un ou l'autre, par exemple, traiter ou ne pas traiter.

Le paradigme de la discrimination vise à cette dichotomisation des patients. Une fois les groupes formés, il n'y a plus aucune distinction entre les individus au sein de chaque groupe, notamment en termes de degré de gravité de la maladie pour les patients supposés malades. Un critère à optimiser peut être le nombre moyen d'erreurs de classement commises par le marqueur, même s'il sera démontré par la suite qu'il n'est pas le plus utile. Si $Y = 1$ dénote le fait que le test associé au marqueur soit positif (i.e. valeur du marqueur au dessus ou en dessous du seuil de positivité) et $Y = 0$ le fait que le test soit négatif, la fonction à minimiser est donnée par :

$$n \times (P(Y = 0, M = 1) + P(Y = 1, M = 0)) \tag{2.2}$$

où $P(Y = 0, M = 1)$ dénote la probabilité conjointe qu'un patient soit malade et que son test soit négatif et $P(Y = 1, M = 0)$ correspond à la probabilité qu'il soit non malade et que son test soit positif.

Calibration et discrimination sont donc deux mesures de performance différentes, qui peuvent conduire à un choix de marqueur différent. Diamond (1992) a en effet montré que l'optimisation de l'une de ces mesures peut aller à l'encontre de l'autre, car elles sont basées sur

des échelles différentes. La discrimination repose sur une échelle catégorielle binaire, alors que la calibration est basée sur une échelle continue. De nombreux articles arguent en faveur de l'une ou l'autre des mesures, en se basant uniquement sur le fait qu'elles ne conduisent pas au même résultat ; il est important de souligner que leurs objectifs sont différents. Mais ni l'une ni l'autre ne tient compte d'un aspect essentiel en analyse de décision, qui est l'utilité attendue de l'action à réaliser.

2.2.4 La prise en compte de l'utilité

La fonction d'optimisation utilisée classiquement en discrimination (équation 2.2) tient compte des deux types d'erreur de classement possibles – faux positifs et faux négatifs – en leur donnant le même poids. En réalité, il est rare que les conséquences de ces deux types d'erreurs soient de la même gravité en termes d'état de santé. Soit un patient de 40 ans présentant une hémiparésie droite modérée et des céphalées ; la réalisation d'un scanner peut permettre de détecter la présence d'une tumeur, opérée par la suite. En moyenne, avec une tumeur non opérée, un patient peut espérer 2,2 années de vie ; une opération lorsque le patient ne présente pas de tumeur conduit en moyenne à 20 années de vie, alors qu'un patient sans tumeur et non opéré peut espérer en moyenne vivre 21 ans. Ici, le fait de ne pas traiter un patient malade est beaucoup plus coûteux que de traiter à tort un patient non malade. En fonction de la problématique de santé, la situation en termes d'état de santé espéré n'est pas aussi tranchée que dans cet exemple. Néanmoins, un marqueur qui minimise le nombre moyen d'erreurs de classement n'est pas forcément celui qui est le plus utile en termes d'état de santé espéré pour la population considérée.

Historiquement, le fait de tenir compte des conséquences attendues des actions qui seront prises d'après des données sur la façon de conduire leur analyse remonte à l'époque des Lumières. Condorcet affirmait déjà que pour condamner une personne, il fallait tenir compte du coût de condamner un innocent par rapport au coût d'innocenter un coupable (Condorcet, 1785). Plus tard, l'utilisation des principes d'optimalité dans l'évaluation des procédures statistiques apparaît dans certains développements clés, comme l'estimation de l'efficacité (Fisher, 1925) et l'analyse de la puissance des tests (Neyman et Pearson, 1933).

Ainsi, le choix d'un marqueur ne doit pas tenir compte uniquement des probabilités d'erreur de classement, mais également des coûts associés à chacun des types d'erreurs et des bénéfices associés à chacun des types de classement corrects. A l'issue de la mesure du marqueur et de la prise de décision, quatre situations sont envisageables : le patient est traité alors

qu'il est malade (situation notée z_{TM}, T pour traité et M pour malade), traité alors qu'il n'est pas malade (notée $z_{T\bar{M}}$, $\bar{M}$ correspondant au fait que le patient n'est pas malade), ou bien le patient peut ne pas être traité alors qu'il est malade ($z_{\bar{T}M}$), ou bien encore non traité et non malade ($z_{\bar{T}\bar{M}}$). Ces situations sont présentées sur l'arbre de décision de la figure 2.1 .

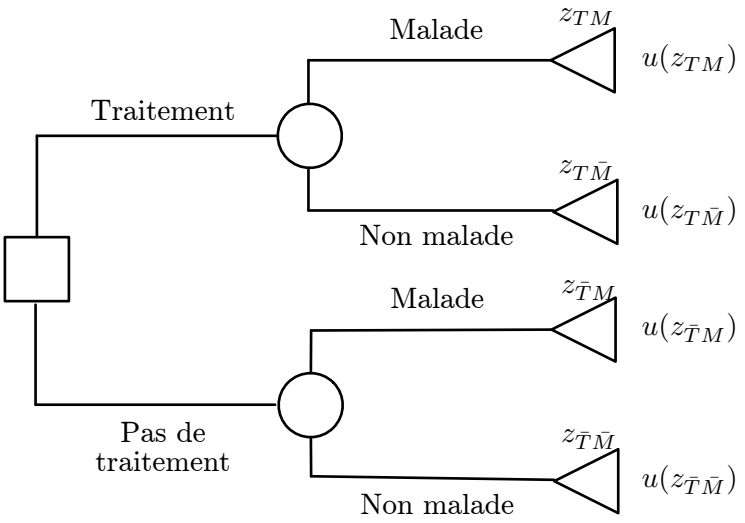

Figure 2.1 – Arbre de décision concernant la décision de traitement, avec les valeurs d'utilité associées à chacune des situations possibles par rapport à la décision prise vis à vis du vrai statut du patient.

A chacune de ces situations peut être associée une *utilité*, notée $u(z)$, correspondant à une mesure de l'état de santé attendu pour chaque situation, par exemple, en termes de qualité de vie, de quantité de vie, ou de qualité de vie ajustée sur la quantité de vie. En règle générale, l'utilité associée à la situation non malade et non traité est la meilleure des quatre utilités et celle associée à la situation malade et non traité la moins bonne. Les utilités correspondant aux deux autres situations possibles se situent entre ces extrêmes, mais leur ordre relatif peut varier.

Plusieurs méthodes ont été décrites afin d'estimer ces utilités à partir d'un patient, ou d'un ensemble de patients, ces méthodes reposant notamment sur la théorie des jeux (Parmigiani, 2002). Un exemple est donné ci-après. Soit $u(z^0)$ l'utilité associée à un état de santé supposé parfait (z^0), et $u(z_0)$ l'utilité associée au plus mauvais état de santé envisageable (z_0), par exemple, le décès. $u(z^0)$ et $u(z_0)$ sont fixées arbitrairement à 1 et 0. Un processus aléatoire conduit à l'état z^0 avec une probabilité α, et à l'état z_0 avec une probabilité $1-\alpha$. L'utilité moyenne de ce processus est donc :

$$\alpha \times u(z^0) + (1-\alpha) \times u(z_0) = \alpha$$

Ce processus est comparé au processus qui conduit de manière certaine à l'état traité et non malade $z_{T\bar{M}}$ (par exemple). L'utilité moyenne de ce second processus est donc de $u(z_{T\bar{M}})$. L'étape suivante consiste à demander à un patient, ou à un ensemble de patients, la valeur de probabilité α pour laquelle ils seraient indifférents entre le choix de l'un ou l'autre des processus pour la suite de leur vie. Les patients étant indifférents entre les deux processus, les utilités moyennes des deux processus sont égales pour cette valeur α. Dans ce cas, $u(z_{T\bar{M}})$ correspond à la valeur de α déterminée par les patients. Cette méthode d'estimation est adaptée lorsque l'utilité est estimée en termes de qualité de vie ; d'autres méthodes ont été proposées lorsque l'utilité est mesurée en termes de quantité de vie, ou en termes de quantité de vie ajustée sur la qualité de vie. Il sera tout de même montré, par la suite, qu'il n'est pas toujours nécessaire, dans le cas du choix d'un marqueur, de quantifier séparément ces quatre utilités.

Soit a le marqueur retenu pour classer les patients et Y^a le résultat du test basé sur ce marqueur. La probabilité de chacune des situations en fonction du marqueur retenu est notée $P_a(z|\theta)$, où θ correspond à un ensemble de paramètres. Une fois ces différents éléments spécifiés, le meilleur marqueur peut être choisi en faisant appel à la théorie de la décision.

2.2.5 Théorie de la décision

La mesure d'un marqueur ne permet pas de connaître de manière certaine le vrai statut du patient. Un clinicien sait juste qu'en moyenne, si le test associé au marqueur est positif, le patient a par exemple 70 % d'être réellement malade ; à l'inverse, si le test est négatif, le patient peut avoir 80 % de ne pas être malade. Le fait que tous les patients ayant un test positif ne soient pas malades et, de même, que tous les patients ayant un test négatif ne soient pas forcément des non malades, est lié à la variabilité des valeurs du marqueur chez les malades et les non malades. Ainsi, le décideur doit prendre une décision concernant le choix du marqueur en tenant compte du caractère incertain du résultat du test en fonction du statut du patient ; c'est une prise de décision dite en situation de risque, en raison de cette incertitude.

Cette situation est très courante dans les paris. A un jeu de pile ou face, il est proposé à une personne de gagner un euro si elle prédit correctement la face sur laquelle la pièce tombe et de perdre un euro et cinquante centimes si sa prédiction est mauvaise. Le parieur ne sait pas à l'avance sur quelle face la pièce va tomber, il sait seulement qu'en moyenne, la pièce a 50 % de chance de tomber sur chacune des faces. Il doit donc prendre la décision de jouer ou de ne pas jouer sans connaître de manière certaine les conséquences de son action ; en revanche, il connaît les coûts, les bénéfices et la probabilité de chacune des situations possibles à l'issue du pari.

Intuitivement, le joueur sait qu'il n'a pas intérêt à parier, car en pariant un grand nombre de fois, il perdra de l'argent. La théorie de la décision permet de formaliser l'intuition du joueur et la prise de décision en situation de risque.

Une décision rationnelle consiste à choisir l'action (i.e. le marqueur) qui conduit, en moyenne, à la meilleure utilité. L'utilité moyenne associée à l'action a est donnée par l'intégrale, sur l'ensemble des situations possibles suite à la prise de décision, de la probabilité de la situation multipliée par l'utilité associée :

$$\mathcal{U}(a) = \int_{\mathcal{Z}} P_a(z|\theta)u(z)\, dz \tag{2.3}$$

où $\mathcal{U}(a)$ correspond à l'utilité moyenne, ou utilité *espérée* si l'action a est retenue.

A l'issue de la décision prise suite à la mesure d'un marqueur, quatre situations sont envisageables ; l'utilité espérée est donnée par :

$$\mathcal{U}(a) = P_a(z_{YM}|\theta)u(z_{YM}) + P_a(z_{\bar{Y}M}|\theta)u(z_{\bar{Y}M}) + P_a(z_{Y\bar{M}}|\theta)u(z_{Y\bar{M}}) + P_a(z_{\bar{Y}M}|\theta)u(z_{\bar{Y}\bar{M}})$$

Cette utilité peut être réécrite en remplaçant θ par les probabilités associées à chacune des situations :

$$\begin{aligned}\mathcal{U}(a) =& P(Y^a = 1, M = 1)u(z_{Y^a,M}) + P(Y^a = 0, M = 1)u(z_{\bar{Y}^a,M}) \\ &+ P(Y^a = 1, M = 0)u(z_{Y^a,\bar{M}}) + P(Y^a = 0, M = 0)u(z_{\bar{Y}^a M})\end{aligned} \tag{2.4}$$

où Y^a correspond au résultat du test basé sur le marqueur a. L'utilité ainsi obtenue correspond à l'utilité que le décideur peut espérer en moyenne s'il utilise le marqueur a dans la population cible. Le meilleur marqueur a_B est celui qui maximise l'utilité espérée :

$$a_B = \arg\max\, \mathcal{U}(a)$$

Au travers de l'utilité, il est donc possible de mesurer la capacité du test à discriminer les patients en tenant compte des coûts associés aux erreurs de classement et des bénéfices associés aux classements corrects. L'utilité peut être calculée à l'échelle individuelle ou populationnelle. A l'échelle populationnelle, les coûts et les bénéfices du traitement sont évalués en moyenne sur la population. A l'échelle individuelle, ces coûts et bénéfices sont déterminés par le patient, en fonction de son aversion vis à vis de la maladie et de ses craintes par rapport au traitement.

Néanmoins, même si l'utilité d'un traitement peut être évaluée pour un individu, elle garde un aspect populationnelle, la probabilité conjointe du résultat du test et de la maladie étant estimée par un groupe de patients. En introduisant des covariables dans la modélisation de cette probabilité, reflétant certaines caractéristiques des patients, il est possible de se rapprocher de plus en plus de l'utilité d'un marqueur pour un individu.

2.2.6 Des outils pour des échelles

Lorsqu'un marqueur est évalué selon sa capacité à discriminer les patients, les différentes situations possibles suite à l'utilisation du marqueur sont directement spécifiées ; il est donc relativement facile d'introduire des notions de coût et de bénéfice dans le choix du marqueur, comme effectué dans l'équation (2.4). Lorsque la calibration d'un marqueur est analysée, les conséquences liées à l'utilisation du marqueur ne sont pas directement spécifiées. Les coûts et les bénéfices interviennent dans ce cas au moment où le clinicien et le patient prennent une décision en fonction de la probabilité de maladie prédite par le marqueur, au travers d'une discussion, et non lors du choix du marqueur.

Plusieurs mesures existent donc pour évaluer un marqueur, certaines intégrant directement les coûts et bénéfices de l'action envisagées, d'autres non ; certaines plus pour objectif d'acquérir la meilleure information possible sur un patient, et se rapprochent plus de l'individu, d'autres ont pour objectif principal de former des groupes de patients, et s'appliquent plus à l'échelle populationnelle. Les réflexions menées en santé publique tentent toutefois de concilier l'intérêt collectif et l'intérêt individuel. De nombreux outils ont été développés pour choisir des marqueurs en fonction d'une mesure de performance. Le chapitre trois présente certains d'entre eux, en essayant de mettre en évidence leurs avantages, leurs limites et surtout leur complémentarité.

2.3 Seuil de positivité

Dans la partie précédente, il a été omis que, pour utiliser un marqueur quantitatif comme test diagnostique ou pronostique, il faut définir un seuil de positivité au dessus ou en dessous duquel le test est jugé positif et le patient considéré comme malade ou allant développer la maladie. Les performances d'un marqueur dépendent donc du seuil de positivité retenu ; la comparaison des différents marqueurs doit être effectuée pour tous les seuils possibles de chacun d'entre eux. Il sera toutefois montré par la suite que, dans certains cas, le choix du marqueur

peut être effectué sans tenir compte des différents seuils possibles. Une fois le marqueur retenu, il reste donc à estimer le seuil optimal, ainsi qu'un intervalle de confiance.

2.3.1 Estimation du seuil optimal et de son intervalle de confiance

Le choix du seuil optimal d'un marqueur se place résolument dans l'objectif de discrimination d'individus selon leur état de santé probable, actuel ou futur. Ce choix peut être effectué sur l'échelle de mesure du marqueur, ou bien sur l'échelle du risque prédit par le marqueur, en retenant un risque au dessus duquel il est utile de traiter le patient. On notera par la suite c le seuil de positivité et $\mathcal{U}(c)$ l'utilité espérée associée au seuil c pour un marqueur donné. Le seuil optimal est celui qui maximise l'utilité espérée.

Dans l'équation (2.4), la probabilité conjointe du résultat du test et du statut du patient peut être décomposée en une probabilité de résultat sachant le statut multipliée par la probabilité de la maladie ($P(Y|M) \times P(M)$). La probabilité que le test soit positif sachant que l'individu est malade est appelée sensibilité ; elle dépend du seuil de positivité retenu et de la distribution du marqueur chez les malades ; elle sera notée $\text{Sen}(c)$. La probabilité que le test soit négatif sachant que l'individu n'est pas malade est appelée spécificité du test (notée $\text{Spe}(c)$) ; elle dépend de la distribution du marqueur chez les non malades et du seuil de positivité. Si π correspond à la prévalence de la maladie dans la population étudiée, l'équation (2.4) peut être réécrite de la façon suivante :

$$\mathcal{U}(c) = \text{Sen}(c)\pi u(z_{YM}) + (1-\text{Sen}(c))\pi u(z_{\bar{Y}M}) + (1-\text{Spe}(c))(1-\pi)u(z_{Y\bar{M}}) + \text{Spe}(c)(1-\pi)u(z_{\bar{Y}\bar{M}})$$

Le seuil qui maximise cette fonction maximise également la fonction suivante :

$$\mathcal{U}^{\star}(c) = \text{Sen}(c) + \text{Spe}(c) \times \frac{(u(z_{\bar{Y}\bar{M}}) - u(z_{Y\bar{M}}))}{(u(z_{YM}) - u(z_{\bar{Y}M}))} \times \frac{(1-\pi)}{\pi} \tag{2.5}$$

$u(z_{\bar{Y}\bar{M}}) - u(z_{Y\bar{M}}) = CN$ correspond au coût de traiter un patient non malade par rapport à ne pas le traiter ; $u(z_{YM}) - u(z_{\bar{Y}M}) = BN$ correspond au bénéfice net de traiter un patient malade par rapport à ne pas le traiter. Le ratio BN/CN est appelé ratio bénéfice net sur coût net. Il peut s'interpréter comme le nombre de patients non malades qu'un clinicien est prêt à traiter à tort pour ne pas manquer le traitement d'un patient réellement malade (DeNeef et Kent, 1993). Plus le ratio est élevé, plus le seuil choisi privilégie la sensibilité du test ; à l'inverse, plus ce ratio est faible, plus le seuil optimal favorise la spécificité du test. Ainsi, il n'est pas nécessaire

de quantifier les utilités associées aux quatre situations possibles, mais uniquement ce ratio, à la signification très concrète.

L'interprétation donnée précédemment à ce ratio s'inscrit dans une démarche d'estimation de l'utilité du traitement à l'échelle populationnelle : les coûts et bénéfices sont estimés en moyenne pour un ensemble de personnes. Néanmoins, des méthodes ont été mentionnées précédemment pour estimer les quatre utilités qui composent ce ratio à l'échelle du patient (partie 2.2.4). Par la suite, ce ratio BN/CN est utilisé sans préciser s'il a valeur pour un ensemble d'individus, ou pour un unique patient, ceci dépendant du contexte.

Le seuil c qui maximise (2.5) est tel que :

$$\frac{d\text{Sen}(c)/dc}{d\text{Spe}(c)/dc} = -\frac{1-\pi}{\pi} \times \frac{CN}{BN}$$

Lorsque la distribution du marqueur suit une loi normale chez les malades et les non malades, il existe une formule explicite du seuil optimal. Si les paramètres de distribution sont estimés par maximum de vraisemblance à partir d'un échantillon de patients malades et non malades, une estimation ponctuelle du seuil est obtenue en remplaçant les paramètres dans la formule explicite par leur estimation du maximum de vraisemblance. Des méthodes similaires ont été proposées pour des marqueurs suivant des lois gamma, log normales ou normales après transformation de Box-Cox (Fluss et *al.*, 2005 ; Schisterman et Perkins, 2007). Ces dernières méthodes ne tiennent pas compte du ratio BN/CN, ni de la prévalence, mais ces deux paramètres pourraient être introduits facilement. Lorsqu'une formule du seuil optimal existe, un intervalle de confiance peut être obtenu grâce à la méthode Delta ; en l'absence de formule explicite, le seuil optimal est calculé par des méthodes numériques ; un intervalle de confiance peut être obtenu par bootstrap. Néanmoins, la probabilité de couverture de l'intervalle de confiance du seuil optimal obtenue par bootstrap est parfois éloignée de celle souhaitée (Schisterman et Perkins, 2007).

Un des objectifs du travail de thèse a été de développer une méthode d'estimation du seuil optimal et de son intervalle de confiance qui soit valable quelle que soit la distribution suivie par le marqueur chez les malades et les non malades. De plus, dans le cas d'un marqueur estimé à partir du profil d'un biomarqueur dynamique, l'intervalle de confiance du seuil optimal doit inclure l'incertitude liée à l'estimation des valeurs du marqueur, ce qui n'est pas encore possible directement avec les méthodes existantes. Cette source d'incertitude a été incluse dans la nouvelle méthode proposée et présentée dans le chapitre cinq.

2.3.2 Seuil optimal et préférences individuelles

Le seuil optimal dépend du ratio bénéfice net sur coût net. A l'échelle collective, d'une équipe de soin à l'autre, les préférences en termes de sensibilité et de spécificité peuvent varier. De même, à l'échelle individuelle, certains patients craignent avant tout les conséquences de la maladie ; cette crainte nécessite une bonne sensibilité du test, alors que d'autres redoutent surtout les conséquences du traitement lui-même ; dans ce cas, la spécificité est à privilégier. Ces variations dans les préférences collectives ou individuelles se traduisent par des variations du ratio bénéfice sur coût net, entraînant ainsi des variations du seuil optimal. Un des derniers objectifs du travail de thèse a été de tenir compte de la variabilité des préférences des patients et des cliniciens dans la détermination du seuil. Cette partie est présentée dans le chapitre six.

Deuxième partie

Estimation et comparaison de marqueurs

Chapitre

3 Comparaison de marqueurs

Le choix d'un marqueur pour le diagnostic, le diagnostic précoce ou le pronostic passe par l'évaluation et la comparaison des performances des différents marqueurs envisageables. Plusieurs mesures de performance ont été introduites dans la première partie : la calibration, la discrimination et l'utilité. A chacune de ces " échelles " de mesure peuvent être associés des " outils " de mesure ; certains sont spécifiques d'une échelle, d'autres essayent d'en intégrer plusieurs au sein d'une même valeur ou d'un même graphique. L'objectif de cette partie n'est pas de présenter tous les outils possibles, mais un certain nombre d'entre eux, en expliquant à quoi ils correspondent, quels sont leurs avantages et leurs limites, ainsi que leur complémentarité.

Dans tous les cas, le choix d'un marqueur repose sur l'optimisation d'une fonction, que ce soit par exemple la fonction du nombre moyen d'erreurs de classement (partie 2.2.3) ou la fonction d'utilité (partie 2.2.5). Selon Murphy et Winkler (1987), toute l'information nécessaire à l'évaluation d'un marqueur est contenue dans la distribution de probabilité conjointe des valeurs de marqueur et du statut du patient vis à vis de la maladie. Le calcul du nombre moyen d'erreurs de classement ne fait en effet intervenir que les distributions de probabilité conjointes (équation (2.2)) ; le calcul de la fonction d'utilité nécessite en plus les utilités associées aux quatre situations possibles (équation (2.4)), mais toute l'information relative au marqueur en lui-même est bien contenue dans la distribution de probabilité conjointe.

Cette distribution conjointe est factorisable de deux façons différentes. La première factorisation possible fait intervenir la distribution des valeurs du marqueur sachant le statut du patient et la distribution marginale des statuts des patients ; la seconde, à l'inverse, repose sur la

distribution des statuts des patients sachant la valeur du marqueur et la distribution marginale des valeurs du marqueur. Différents outils de mesure de performance existent en fonction de la factorisation envisagée. La première factorisation sera qualifiée par la suite de " factorisation orientée marqueur " et la seconde de " factorisation orientée patient ".

De manière générale, on notera f_0 et f_1 les densités de probabilité des valeurs de marqueurs chez les malades et les non malades et F_0 et F_1 les fonctions de répartition.

3.1 La factorisation orientée marqueur

Pour la factorisation orientée marqueur, la distribution de probabilité conjointe du résultat du test et du statut du patient est factorisée en la probabilité du résultat du test sachant le statut du patient multipliée par la probabilité du statut du patient ($P(Y|M) \times P(M)$). $P(M)$ représente la probabilité de maladie prédite avant la mesure du marqueur ; elle est obtenue à partir de la prévalence de la maladie dans la population. La probabilité du résultat du test sachant le statut du patient représente l'information supplémentaire apportée par le marqueur (Murphy et Winkler, 1987).

3.1.1 Cas d'un marqueur binaire

Lorsqu'un marqueur ne prend que des valeurs binaires (0 pour les patients supposés non malades et 1 pour ceux supposés malades), les performances peuvent être évaluées à l'aide de la sensibilité et de la spécificité. La sensibilité est la probabilité que le test soit positif sachant que le patient est malade ($\text{Sen} = P(Y = 1|M = 1)$). Appelée aussi proportion de vrais positifs par opposition à la proportion de faux négatifs ($1-\text{Sen}$), la sensibilité mesure la capacité du marqueur à détecter les malades. La spécificité est la probabilité que le test soit négatif sachant que le patient n'est pas malade ($\text{Spe} = P(Y = 0|M = 0)$). Appelée aussi proportion de vrais négatifs, par opposition à celle de faux positifs ($1 - \text{Spe}$), la spécificité mesure la capacité du marqueur à identifier les non malades. Plus ces deux indices sont élevés, meilleur est le test associé au marqueur. La sensibilité et la spécificité ne dépendent pas de la prévalence de la maladie dans la population étudiée, mais des caractéristiques de la population. Ainsi, il est important, entre autres, que la population de malades étudiée reflète l'ensemble du panorama de la maladie pour une bonne évaluation du marqueur. Une étude n'incluant, parmi les patients malades, que des patients dont la gravité de la maladie est élevée pourrait surestimer la sensibilité du marqueur,

cette dernière étant souvent plus élevée chez ces patients que chez ceux dont la gravité de la maladie est légère à modérée.

Si deux marqueurs conduisent à des valeurs de spécificité similaires mais à des sensibilités différentes, le marqueur ayant la meilleure sensibilité est à privilégier et vice versa. Lorsqu'un marqueur a une bonne sensibilité mais une spécificité médiocre et qu'un second marqueur présente une bonne spécificité mais une sensibilité inférieure à celle du premier, le choix du marqueur n'est pas réalisable de manière automatique. Il faut, dans ce cas, définir une priorité concernant le type d'erreurs de classement à limiter – faux positifs ou faux négatifs – et le choix du marqueur dépend par conséquent du contexte.

3.1.2 Cas d'un marqueur quantitatif

Lorsqu'un marqueur prend des valeurs quantitatives, il faut définir un seuil – noté c – au dessus ou en dessous duquel le test associé est défini comme positif. Par convention, on considérera par la suite que les valeurs élevées du marqueur sont en faveur de la maladie. Dans ce cas, le test est dit positif lorsque la valeur du marqueur est supérieure au seuil. De manière similaire, le test pourrait être défini comme positif lorsque la valeur du marqueur est inférieure au seuil de positivité dans le cas où les valeurs faibles du marqueur sont en faveur de la maladie. La sensibilité et la spécificité peuvent être calculées pour toutes les valeurs de seuil de positivité ; on notera $\mathrm{Sen}(c)$ la sensibilité associée au seuil de positivité c, ($\mathrm{Sen}(c) = P(Y > c|M = 1)$) et $\mathrm{Spe}(c)$ la spécificité associée à ce seuil ($\mathrm{Spe}(c) = P(Y \leq c|M = 0)$).

Les mesures de performance sur l'ensemble des seuils de positivité possibles sont souvent résumées au travers d'une courbe ROC, pour Receiver Operating Characteristic. Ces courbes sont apparues durant la seconde guerre mondiale, suite à l'attaque de Pearl Harbor, afin d'affiner la capacité à détecter des avions ennemis à l'aide d'un signal radar. Elles ont été introduites dans la littérature médicale par Hanley et McNeil (Hanley et McNeil, 1982). Une courbe ROC correspond au tracé de la sensibilité en fonction du complément de la spécificité pour l'ensemble des seuils de positivité possibles d'un marqueur ; la figure 3.1 représente la courbe associée à un marqueur quantitatif ne prenant que neuf valeurs différentes.

Une courbe ROC est contenue dans un carré de côté un. La première bissectrice du graphique correspond à un marqueur qui ne classerait pas mieux les patients que le simple hasard ; ainsi, un bon marqueur est un marqueur dont la courbe s'éloigne le plus possible de la première bissectrice et se rapproche du coin supérieur gauche du graphique. L'avantage de la courbe ROC

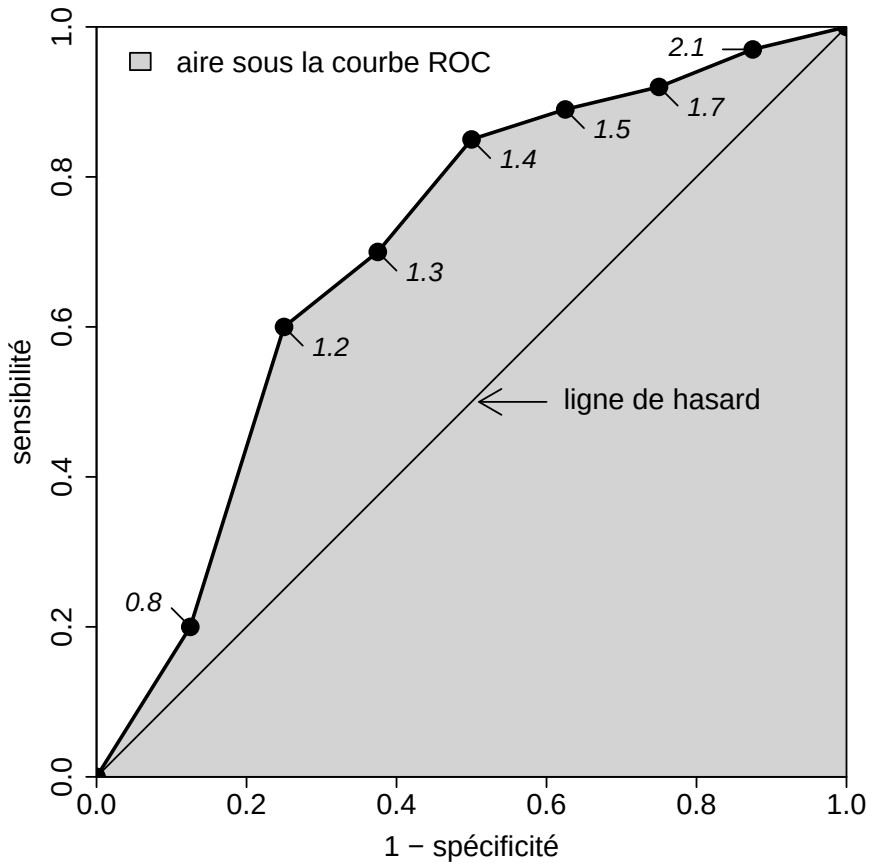

Figure 3.1 – Courbe ROC associée à un marqueur quantitatif ne prenant que neuf valeurs différentes. Les valeurs indiquées à côté des points de la courbe ROC correspondent aux valeurs de marqueur conduisant à de telles sensibilités et spécificités.

est qu'elle résume la capacité globale du marqueur à discriminer les patients sur l'ensemble des seuils de positivité possibles. Elle permet facilement de comparer des marqueurs, même lorsqu'ils ont des échelles de mesure très différentes.

Soit deux marqueurs dont les courbes ROC sont représentées sur la figure 3.2. Le marqueur 1 a une courbe ROC toujours supérieure à celle du marqueur 2. Ceci implique que, pour un seuil du marqueur 1 et un seuil du marqueur 2 tels que les spécificités soient identiques, la sensibilité correspondante du marqueur 1 est supérieure à celle du marqueur 2 ; ceci est valable quel que soit le niveau de spécificité considéré. De même, pour une sensibilité identique pour les deux marqueurs, la spécificité correspondante du marqueur 1 est supérieure à celle du marqueur 2, et ce, quel que soit le niveau de sensibilité considéré. Ainsi, quels que soient les niveaux de sensibilité ou de spécificité désirés, le marqueur 1 est à privilégier.

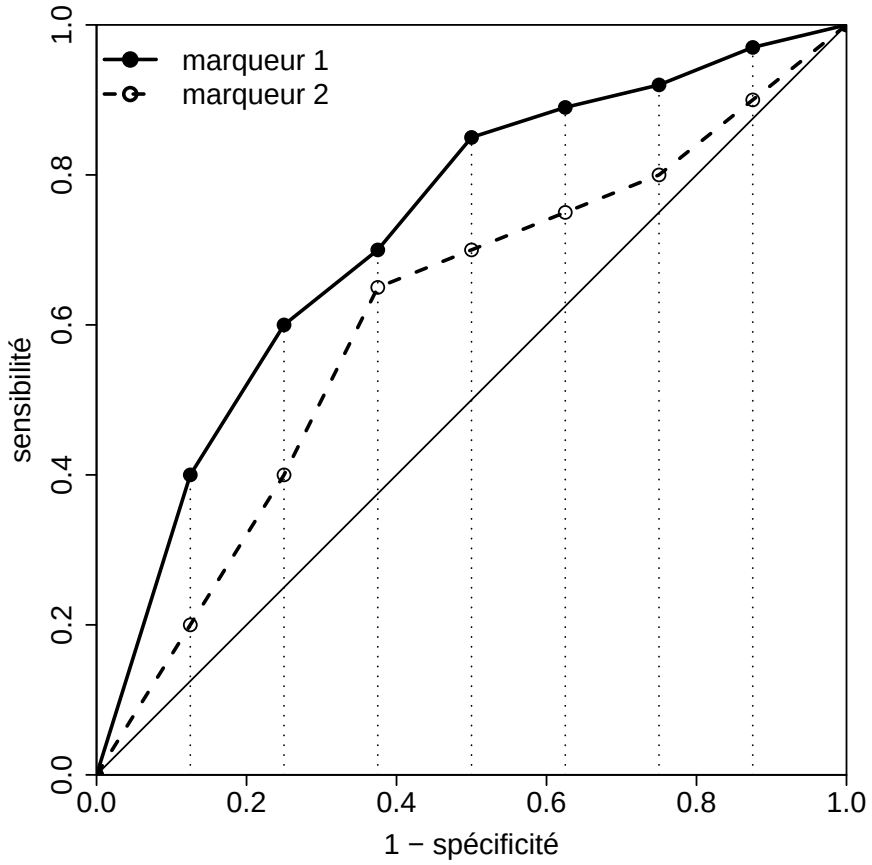

Figure 3.2 – Courbes ROC associées à deux marqueurs quantitatifs.

L'aire sous la courbe ROC (notée par la suite AROC) est une mesure quantitative de la performance globale d'un marqueur, sur l'ensemble des seuils de positivité possibles. Mathématiquement, elle est définie par :

$$AROC = \int_0^1 \mathrm{Sen}((1-\mathrm{Spe})^{-1}(t))\, dt \tag{3.1}$$

L'AROC est comprise entre 0 et 1 ; plus elle est élevée, meilleur est le marqueur. Dans le cas de l'exemple précédent, l'AROC du premier marqueur est de 0,73 et celle du second de 0,61 ; il faut donc privilégier le premier marqueur. Une AROC de 0,5 correspond à un marqueur qui ne classe pas mieux les individus que le simple hasard ; un bon marqueur a par conséquent une AROC bien supérieure à 0,5. De par sa définition, l'AROC correspond à la moyenne de la sensibilité pour toutes les valeurs possibles de proportion de faux positif. Elle peut également s'interpréter comme la probabilité qu'une valeur de marqueur d'un patient malade (Y_1) soit supérieure à celle

d'un patient non malade (Y_0) :

$$\begin{aligned} AROC &= \int_0^1 \text{Sen}((1-\text{Spe})^{-1}(t))\, dt \\ &= \int_{-\infty}^{\infty} \text{Sen}(c)\, d(1-\text{Spe}(c)) \\ &= \int_{-\infty}^{\infty} P(Y_1 > c) f_0(c)\, dc \\ &= \int_{-\infty}^{\infty} P(Y_1 > c, Y_0 = c)\, dc \\ &= P(Y_1 > Y_0) \end{aligned} \tag{3.2}$$

D'après ce dernier résultat, l'AROC ne dépend pas des valeurs réelles du marqueur, mais de leur ordre relatif entre les malades et les non malades. C'est une statistique de rang, invariante par transformation monotone croissante des valeurs de marqueur. Ceci se reflète d'ailleurs dans les méthodes d'estimation et de comparaison des AROC, l'AROC étant tout simplement estimée à partir de la statistique de rangs de Mann et Whitney (Pepe, 2003) :

$$AROC = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(y_{1_i} > y_{0j}) + 1/2 \times I(y_{1i} = y_{0j})}{n_1 \times n_0}$$

n_1 correspond au nombre de patients malades, n_0 au nombre de non malades ; y_{1i} correspond à la valeur du marqueur du i$^{\text{ème}}$ individu du groupe malade, et y_{0j} à la celle du j$^{\text{ème}}$ individu du groupe non malade. L'AROC est aussi appelée statistique de c. Du fait qu'elle ne dépend que des rangs des valeurs, la factorisation orientée marqueur mesure donc la capacité du marqueur à discriminer des personnes indépendamment des " étiquettes " de valeurs associées aux mesures du biomarqueur (Murphy et Winkler, 1987).

Une mise en garde est nécessaire : les valeurs d'AROC de deux marqueurs ne sont comparables directement que si les courbes ROC associées sont emboîtées. En effet, dans le cas contraire, même si l'AROC du premier marqueur est supérieure à celle du second, le premier marqueur n'est pas forcément le plus utile, ceci dépend des valeurs de sensibilité ou de spécificité souhaitées. La figure 3.3 représente les courbes ROC de deux marqueurs, le premier ayant une AROC de 0,70 et le second de 0,675. Pour les valeurs de marqueur telles que la spécificité soit supérieure à 0,75 (zone en gris clair), le marqueur 2 conduit à de meilleures sensibilités que le marqueur 1. Si une spécificité supérieure à 0,75 est souhaitée, alors le marqueur 2 est à privilégier, même si son AROC est inférieure à celle du premier. Ainsi, lorsque deux courbes ROC

ne sont pas emboîtées, le choix d'un marqueur doit se faire en fonction des valeurs de sensibilités ou de spécificités souhaitées et non uniquement à partir des AROC.

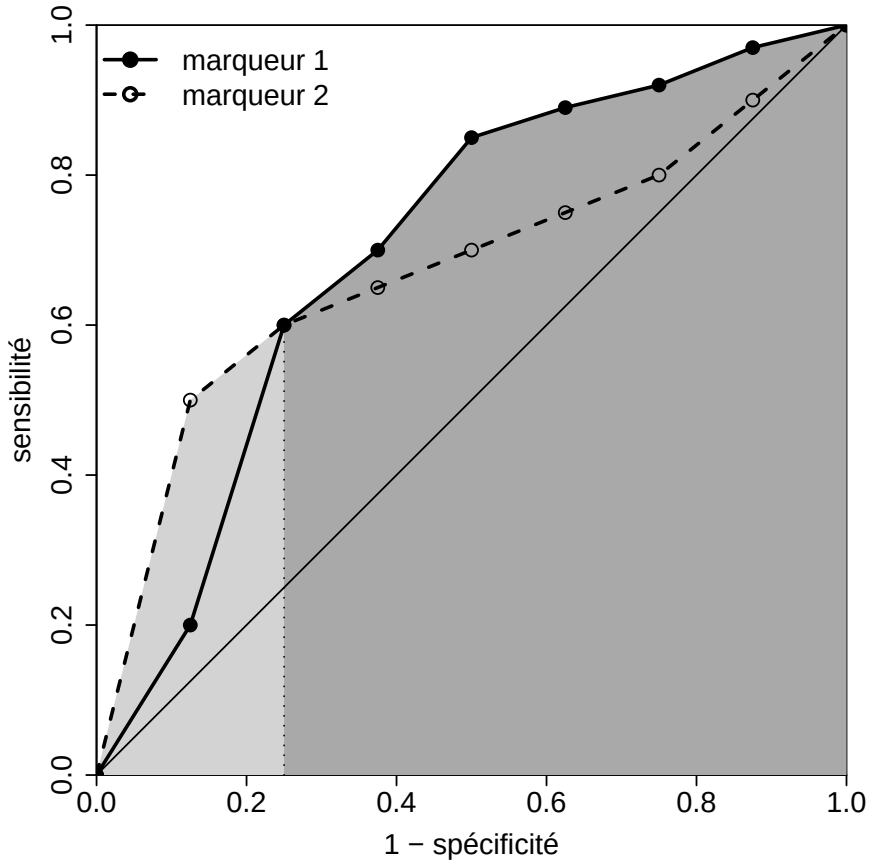

Figure 3.3 – Courbes ROC associées à deux autres marqueurs quantitatifs. La zone en gris foncé est la zone pour laquelle la sensibilité du marqueur 1 est plus élevée que celle du marqueur 2, pour une même valeur de spécificité.

3.1.3 Les limites de l'AROC pour le patient et le clinicien

Pour un patient dont le test associé à la mesure du marqueur est positif, le fait de savoir que la sensibilité du test est de 80 % ne l'aide pas forcément à prendre une décision vis à vis de sa situation. Ce qu'il souhaiterait connaître – lui ou son médecin – est la probabilité qu'il soit malade sachant que le test est positif ; ceci correspond à la valeur prédictive positive, dont l'interprétation est détaillée plus amplement dans la partie suivante ; c'est en fonction de cette probabilité qu'une décision de traitement est prise. De même, si l'AROC représente la probabilité que la valeur d'un marqueur chez un malade soit supérieure à celle d'un non malade, il est rare qu'un clinicien compare deux patients en même temps, en sachant que l'un est malade et l'autre

ne l'est pas. Ainsi, ces mesures de performance semblent ne pas être utiles directement pour le clinicien et le patient.

3.2 La factorisation orientée patient

La factorisation orientée patient décompose la distribution de probabilité conjointe du résultat du marqueur et du statut du patient en la probabilité de maladie sachant la valeur du marqueur, c'est à dire le risque de maladie, multipliée par la probabilité de la valeur du marqueur dans la population ($P(M|Y) \times P(Y)$). Pour cette factorisation, l'intérêt porte sur le risque de maladie prédit en fonction de la valeur du marqueur.

3.2.1 Cas d'un marqueur binaire

Dans le cas d'un marqueur binaire, les performances sont calculables en termes de valeurs prédictives. La valeur prédictive positive, notée Vpp, correspond à la probabilité qu'un individu soit malade sachant que le test associé au marqueur est positif ($\text{Vpp} = P(M = 1|Y = 1)$). La valeur prédictive négative, notée Vpn, correspond à la probabilité qu'un individu soit non malade sachant que le test associé au marqueur est négatif ($\text{Vpn} = P(M = 0|Y = 0)$). Ces deux indicateurs intéressent directement le clinicien et le patient, qui peuvent prendre une décision à partir de la probabilité de maladie prédite par le résultat du test.

Les valeurs prédictives dépendent des caractéristiques de la population et surtout de la prévalence. Ainsi, les résultats obtenus d'une population à l'autre ne sont pas comparables si la prévalence n'est pas identique. Toutefois, si la sensibilité et la spécificité du test ont été évaluées dans une population où la prévalence est de π_1 et si la prévalence dans une autre population est connue (π_2), alors les valeurs prédictives attendues dans la seconde population sont calculables à l'aide du théorème de Bayes :

$$\text{Vpp}_2 = \frac{P(Y = 1|M = 1)\pi_2}{P(Y = 1|M = 1)\pi_2 + P(Y = 1|M = 0)(1 - \pi_2)}$$
$$\text{Vpn}_2 = \frac{P(Y = 0|M = 0)(1 - \pi_2)}{P(Y = 0|M = 1)\pi_2 + P(Y = 0|M = 0)(1 - \pi_2)}$$

3.2.2 Cas d'un marqueur quantitatif

Lorsque le marqueur prend des valeurs continues, les valeurs prédictives peuvent être calculées pour les différentes valeurs de seuil de positivité du marqueur, par exemple, $\text{Vpp}(c) =$

$P(M = 1|Y > c)$. D'après l'équation (2.1) :

$$\frac{\text{Vpp}(c)}{1 - \text{Vpp}(c)} = \underbrace{\frac{P(Y > c|M = 1)}{P(Y > c|M = 0)}}_{(a)} \times \frac{P(M = 1)}{P(M = 0)} = \frac{\text{Sen}(c)}{1 - \text{Spe}(c)} \times \frac{P(M = 1)}{P(M = 0)}$$

De même :

$$\frac{1 - \text{Vpn}(c)}{\text{Vpn}(c)} = \underbrace{\frac{P(Y \leq c|M = 1)}{P(Y \leq c|M = 0)}}_{(b)} \times \frac{P(M = 1)}{P(M = 0)} = \frac{1 - \text{Sen}(c)}{\text{Spe}(c)} \times \frac{P(M = 1)}{P(M = 0)}$$

Les termes (a) et (b) correspondent aux ratios de vraisemblances positifs et négatifs ; c'est au travers de ces ratios que le marqueur apporte de l'information sur le risque de maladie en plus de l'information issue de la prévalence. Plus le ratio de vraisemblance positif est élevé, plus le fait que le marqueur soit supérieur au seuil c accroît la certitude du médecin que le patient est malade, en plus de la simple information contenue dans la prévalence. A l'inverse, plus le ratio de vraisemblance négatif est faible, plus le fait que le marqueur soit inférieur au seuil c accroît la certitude du médecin que le patient n'est pas malade. Le marqueur apporte de l'information si la probabilité de ses valeurs est très différente dans les groupes malades et non malades ; l'information est donc contenue dans la probabilité des valeurs du marqueur sachant le statut du patient, donc d'une certaine façon, dans la sensibilité et la spécificité.

Si la courbe ROC est la généralisation des notions de sensibilité et de spécificité aux cas de marqueurs quantitatifs, il n'existe pas vraiment de généralisation des valeurs prédictives aux cas de marqueurs prenant des valeurs quantitatives. Moskowitz et Pepe (2004) ont proposé de construire une courbe représentant le risque de maladie en fonction des quantiles de la distribution du marqueur chez les malades ; une courbe similaire peut être construite chez les non malades, mais il est difficile de comparer plusieurs marqueurs à partir de deux graphiques séparés.

Lorsque les performances du marqueur sont analysées en termes de risque, une première approche peut simplement consister à comparer les risques relatifs ou les rapports de cotes des marqueurs. Le risque relatif correspond au ratio des risques pour une augmentation d'une unité du marqueur ; plus il est supérieur à 1, plus une augmentation de la valeur du marqueur augmente fortement le risque de maladie. L'estimation du risque de maladie nécessite que l'étude réalisée soit prospective. Une cote (notée Cote) correspond au rapport de la probabilité de maladie sur la probabilité de non maladie pour un niveau donné du marqueur, le rapport de cotes (noté

RC) étant le ratio des cotes pour une augmentation d'une unité du marqueur. Plus le rapport de cotes est élevé, plus l'augmentation des chances de maladie pour une augmentation d'une unité du marqueur est importante. Il peut être estimé à partir d'une étude de cohorte ou d'une étude cas témoins. Rapport de cotes et risque relatif sont des mesures très souvent utilisées en recherche clinique et en épidémiologie.

De nombreuses études se servent, à tort, de l'estimation du rapport de cotes comme d'une mesure de la capacité du marqueur à discriminer les patients. Le rapport de cotes est une mesure d'association et non une mesure de discrimination. S'il est vrai que, plus le rapport de cotes est élevé, meilleure est l'AROC associée, des rapports de cotes jugés élevés ne sont pas forcément associés à des AROC élevées. La figure 3.4 représente les densités de probabilité des valeurs de marqueurs chez les malades et les non malades lorsque celles-ci suivent des lois normales d'écart-type 0,5 dans les deux groupes et de moyenne 0 chez les non malades. La moyenne chez les malades dépend du rapport de cotes choisi pour le marqueur. Sont également indiquées les valeurs d'AROC associées. Même pour des rapports de cotes de 2 ou 3, jugés élevés dans le cas de certaines maladies, le recoupement des densités de probabilité chez les malades et les non malades reste important, ce qui se traduit par des valeurs d'AROC relativement peu élevées. Pour un écart-type de 0,5, il faut un rapport de cotes de plus de 60 – très élevé – afin d'obtenir une valeur d'AROC de 0,93, très satisfaisante.

Cet exemple semble montrer une discordance entre l'AROC et le rapport de cotes. En réalité, lorsque ces critères sont utilisés pour comparer deux marqueurs, les conclusions sont toujours identiques : l'AROC et le rapport de cotes sont plus élevés pour le marqueur qui discrimine le mieux les patients en deux groupes. Ceci est lié au fait que ces deux critères sont calculés à partir de la même information, correspondant aux distributions de probabilité des valeurs de marqueurs chez les malades et les non malades. Pour autant, des valeurs de rapport de cotes relativement élevées ne sont pas forcément associées à des AROC élevées, ce qui crée la confusion.

Le fait qu'un rapport de cotes soit jugé pertinent dépend en réalité du contexte. Dans l'exemple précédent, pour un rapport de cotes de 3, si la prévalence est de 1 %, le risque de maladie lorsque le marqueur vaut 0 est de 0,8 % et passe à 2,5 % pour une valeur de marqueur de 1. Lorsque la prévalence est de 10 %, le risque de maladie lorsque le marqueur vaut 0 est de 8 % et de 21 % lorsque le marqueur vaut 1. Ces deux scénarios peuvent avoir des conséquences très différentes si un risque en dessous de 10 % est considéré comme négligeable et un risque

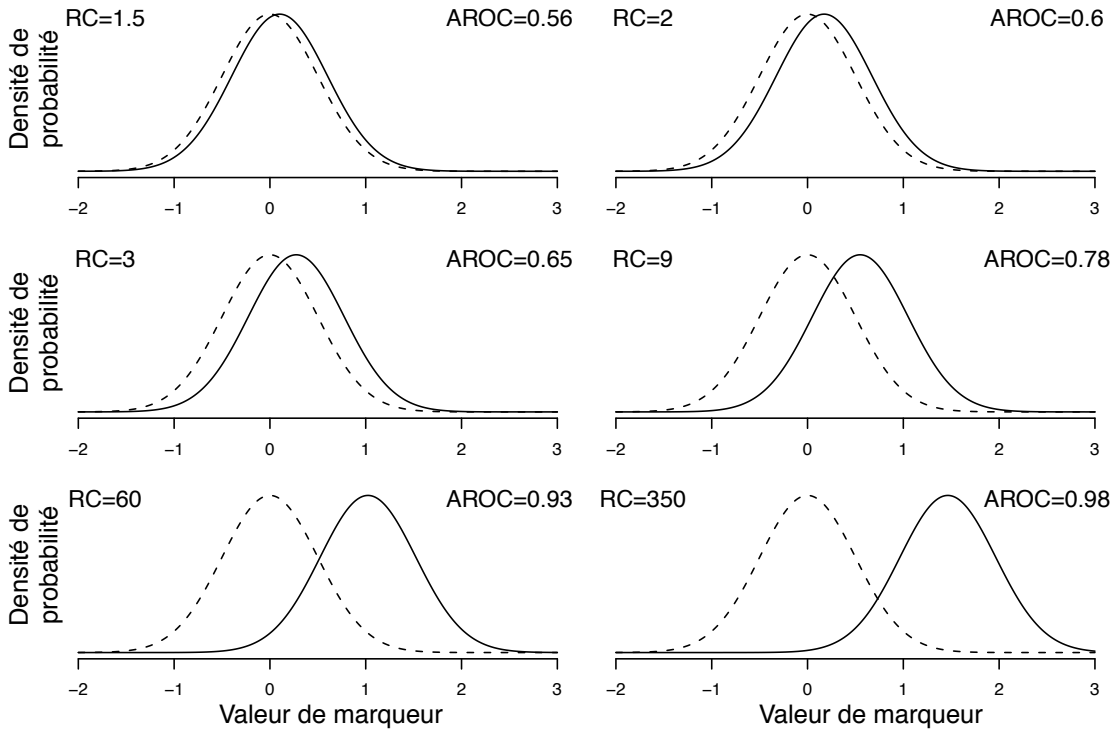

Figure 3.4 – Densité de probabilité des valeurs de marqueurs chez les non malades (pointillés) et chez les malades (traits pleins) associées à différents rapports de cotes, lorsque le marqueur suit des lois normales de variances égales dans les deux groupes.

au dessus de 10 % non négligeable. En termes d'AROC, les deux scénarios sont identiques, car l'AROC ne dépend pas de la prévalence ; dans les deux cas, la capacité du marqueur à discriminer les patients est moyenne (AROC de 0,65). En termes de risque, dans le cas d'une prévalence de 10 %, le marqueur peut avoir un intérêt non pas pour classer les patients en deux groupes, mais pour identifier les patients les plus à risque dans une population et ne proposer un suivi renforcé qu'à une partie de la population. Pour Huang et *al.* (2007), de nombreux marqueurs seraient rejetés s'ils n'étaient évalués qu'au travers de l'AROC, alors qu'ils peuvent présenter un intérêt si l'objectif est d'identifier une partie de la population la plus à risque.

En réalité, le fait de choisir une valeur au dessus de laquelle le risque est jugé pertinent revient indirectement à donner un coût aux faux positifs et faux négatifs et un bénéfice aux vrais positifs et vrais négatifs. Une valeur seuil de risque de 10 % indique que le coût des faux positifs est négligeable devant le bénéfice associé aux vrais positifs, car beaucoup de personnes subiront des examens complémentaires pour rien. La courbe ROC donne le même poids à la sensibilité et à la spécificité et suppose donc des coûts et bénéfices identiques ; la valeur de l'AROC ne doit donc pas être analysée de manière absolue, mais relativement aux bénéfices et aux coûts de

l'action réalisée suite à la mesure du marqueur. Il en va de même pour les valeurs prédites de risque.

Lorsqu'une action bien définie est envisagée suite à la mesure du marqueur, l'objectif est de discriminer les patients. La courbe ROC est un outil utile dans ce cas. Le fait de vouloir sélectionner une partie de la population à un niveau de risque donné – en faisant appel au rapport de cotes ou au risque relatif et à la prévalence – revient également à faire de la discrimination, mais en introduisant indirectement des bénéfices et des coûts attendus si l'action est réalisée via la valeur seuil au delà de laquelle le risque est jugé significatif. Définir une valeur limite de risque est jugé parfois plus simple que de définir explicitement des coûts et des bénéfices ; c'est pourquoi les méthodes reliées à l'analyse du risque sont plus souvent utilisées dans ce cas. Il sera montré par la suite qu'il est également possible d'introduire l'utilité attendue de l'action dans les méthodes reliées à l'analyse de la courbe ROC, sans pour autant définir de manière explicite les coûts et les bénéfices.

Lorsqu'aucune action précise n'est envisagée à l'avance, suite à la mesure du marqueur, mais qu'un médecin souhaite améliorer sa connaissance sur l'état de santé d'un patient pour pouvoir prendre une décision, les méthodes liées à l'analyse du risque sont très appropriées pour choisir un marqueur. Il ne faut pas analyser les risques relatifs ou les rapports de cotes, mais les valeurs absolues de risque prédit, qui tiennent compte de la prévalence. L'objectif est alors que le risque prédit pour un niveau donné du marqueur corresponde bien au risque réel du patient. Ceci revient à analyser la calibration du marqueur et non sa capacité à discriminer les patients.

3.2.3 La calibration et ses limites

La calibration est une mesure de l'adéquation entre le risque prédit pour un niveau du marqueur et le vrai risque des patients à ce niveau (Hand, 1997). La calibration est importante, car en fonction des coûts et des bénéfices de l'action envisagée suite à la mesure du marqueur, le risque de maladie prédit permet de prendre une décision. Pour un risque prédit de 20 %, si le bénéfice attendu à être traité par rapport à ne pas être traité lorsque le patient est réellement malade est très petit par rapport au coût du traitement à tort d'un patient non malade, alors il vaut mieux ne pas être traité. Si le bénéfice attendu par rapport aux coûts est très élevé, il vaut mieux être traité. Ainsi, si le risque prédit par le marqueur ne correspond pas au vrai risque du patient, la décision prise risque d'être éronnée, d'où l'utilité de la calibration.

Le vrai risque à un niveau donné du marqueur est inconnu, mais l'échantillon de population considéré en est représentatif. Une solution souvent adoptée pour mesurer l'adéquation consiste

à calculer des statistiques résumées sur les risques prédits et à les comparer aux statistiques résumées obtenues sur l'échantillon de la population.

La statistique de Hosmer-Lemeshow est une mesure courante de calibration (Hosmer et Lemeshow, 1989). Elle consiste à grouper les patients selon leur risque prédit et à comparer le risque prédit moyen par groupe au risque réel observé au sein de ces groupes. En cas d'adéquation, le résultat suit un χ^2 avec autant de degrés de libertés que de groupes moins 2. La statistique de Hosmer-Lemeshow est une statistique globale sur l'ensemble des niveaux de risque. Il est également possible de représenter les risques prédits en fonction des risques observés par groupe sur un graphique, afin d'analyser la calibration par zones de risque. La courbe associée à un marqueur bien calibré correspond à une droite passant par l'origine et de pente 1.

L'inconvénient de la calibration est qu'elle peut être parfaite sans avoir d'utilité. La règle consistant à affecter comme risque à tous les patients la valeur de la prévalence est parfaitement calibrée, mais elle n'est d'aucune utilité pour prendre une décision en fonction du risque prédit. Si la courbe ROC est peut-être trop contraignante en ne permettant que de classer les patients en deux groupes bien distincts, les mesures de calibration peuvent aboutir à l'extrémité opposée. Elles sont toutefois utiles dans d'autres domaines lorsque, par exemple, la construction d'un modèle de risque incluant plusieurs facteurs a pour but de mieux comprendre certains phénomènes, notamment l'effet des facteurs sur le risque. Dans ce cas, il est important que le risque prédit corresponde au vrai risque. Mais les mesures de calibration " pures " sont d'un intérêt limité dans l'évaluation des performances diagnostiques ou pronostiques d'un marqueur.

Il est tout de même à noter que d'autres mesures de calibration que celle d'Hosmer-Lemeshow existent. Des solutions ont notamment été proposées pour tenir compte du fait que cette statistique peut grouper ensemble des individus dont le vrai risque diffère d'après le profil de covariables, mais dont le risque prédit est de manière erronée identique (Lin et *al.*, 2002 ; Pulkstenis et Robinson, 2002).

3.3 Intégrer discrimination et calibration

Si les mesures de calibration " pures " sont d'un intérêt limité pour la comparaison de marqueurs, plusieurs mesures de calibration intégrant des notions de discrimination ont été proposées, comme par exemple le score de Breier ou les taux de reclassification.

3.3.1 Le score de Breier et l'exactitude

Soit un marqueur prenant des valeurs continues et pour lequel le seuil c a été retenu comme seuil de positivité. Un des inconvénients de la courbe ROC est qu'elle donne la même importance à des erreurs de classement qui n'ont pas la même ampleur : un faux négatif peut être aussi bien un patient dont la valeur du marqueur est juste en dessous du seuil de positivité qu'un patient dont la valeur du marqueur est très en dessous du seuil ; le second cas est tout de même plus problématique.

Une méthode pour introduire la notion d'éloignement entre le vrai statut du patient et le risque prédit par le marqueur consiste à calculer le carré de l'écart entre ces deux valeurs ; ceci correspond au score de Breier. Si M_i dénote le vrai statut d'un patient i, alors le score est donné par :

$$BS = \mathrm{E}((M_i - \hat{P}(M = 1|Y_i))^2)$$

où $\hat{P}(M = 1|Y_i)$ correspond au risque de maladie prédit pour une valeur de marqueur Y_i. Le score de Breier s'estime par la somme des carrés des écarts sur l'ensemble des patients :

$$\hat{BS} = \frac{1}{n}\sum_{i=1}^{n}(M_i - \hat{P}(M|Y_i))^2$$

C'est une mesure d'*exactitude*, c'est à dire une mesure de distance entre le vrai statut du patient et le risque prédit (Hand, 1997). Plus le score de Breier est faible, plus le risque prédit par le marqueur est adéquat, les risques prédits très éloignés du vrai statut étant plus pénalisés que les risques prédits qui diffèrent peu du vrai statut. Le score de Breier peut se décomposer en une somme de deux termes :

- le premier correspond à l'écart entre le risque prédit par le marqueur et le vrai risque du patient ; appelé terme d'*imprécision* par Hand, il mesure la calibration du marqueur ;
- le second terme correspond à l'écart entre le vrai statut du patient et le vrai risque ; appelé par Hand terme d'*inséparabilité*, il mesure la capacité du marqueur à discriminer les patients suivant leur statut.

Ainsi, le score de Breier reflète dans une même mesure la calibration et la capacité à discriminer d'un marqueur. L'inconvénient est de ne pas pouvoir distinguer ces deux propriétés, car suivant les situations, l'une ou l'autre est à privilégier. Enfin, il faut mentionner que d'autres mesures d'exactitude existent, comme le score logarithmique (Hand, 1997).

3.3.2 Les taux de reclassification

Cook (2007) a proposé une autre mesure permettant d'intégrer la discrimination et la calibration : les taux de reclassification, consistant à comparer deux marqueurs au travers du pourcentage de patients qui changent de groupes de risque en utilisant l'un des marqueurs plutôt que l'autre. Cette méthode nécessite de définir à l'avance des groupes de risques, puis de calculer les pourcentages d'individus appartenant à ces groupes d'après le risque prédit par un premier marqueur et enfin de calculer le pourcentage de personnes dont le groupe de risque change si le risque est prédit à partir d'un second marqueur. Ici, aucune distinction n'est effectuée entre les individus dont le risque augmente et qui vont développer la maladie et ceux dont le risque augmente mais qui ne vont pas pour autant développer la maladie ; il en est de même pour les individus dont le risque diminue. Un taux d'amélioration de la classification a été proposé par Pencina et *al.* (2008), qui compte de manière positive les reclassifications dans la bonne direction et de manière négative celles qui sont dans la mauvaise direction, en donnant le même poids, en valeur absolue, aux bonnes et mauvaises reclassifications.

L'objectif de telles mesures n'est pas uniquement que le risque prédit corresponde au risque observé, mais également que les risques des personnes allant développer la maladie et de celles qui ne la développeront pas soient poussés vers les extrêmes. Sans vouloir créer nettement deux groupes, ces mesures favorisent les marqueurs grâce auxquels le maximum de personnes allant développer la maladie seront classées dans des catégories de risque élevé et le maximum de personnes n'allant pas développer la maladie seront classées dans des catégories de risque faible. L'inconvénient de ces mesures est qu'il faut définir à l'avance des niveaux de risques pour former des groupes.

3.3.3 Les courbes de prédiction

Un équivalent continu des taux de reclassification, qui ne nécessite pas la définition de groupes de risque, est la " courbe de prédiction " (traduction peu fiable de " predictiveness curve ") proposée par Pepe et *al.* (2008) et Huang et *al.* (2007). Cette courbe représente le risque de maladie prédit en fonction du quantile de la valeur du marqueur dans la population étudiée. Si $R(\nu)$ correspond au risque associé au $\nu^{\text{ième}}$ quantile du marqueur, alors :

$$R(\nu) = P(M = 1|Y = F^{-1}(\nu))$$

où F est la fonction de répartition du marqueur en question. La figure 3.5 représente les courbes de prédiction associées à deux marqueurs, ainsi que leurs courbes ROC respectives.

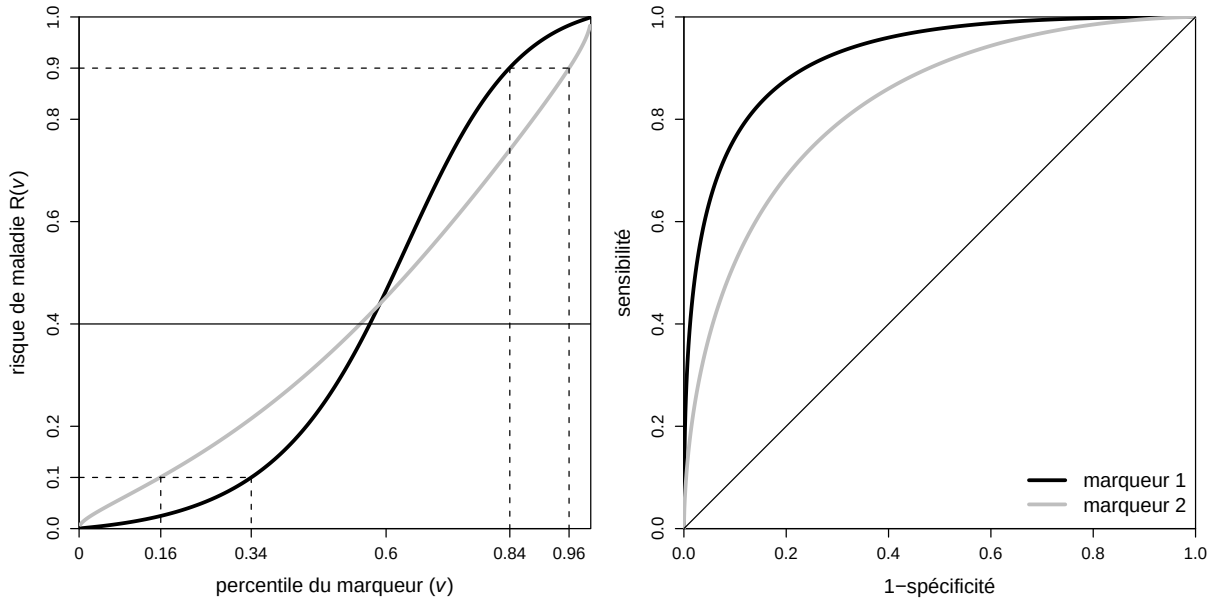

Figure 3.5 – Courbes de prédiction associées à deux marqueurs (à gauche) et courbes ROC (à droite).

A un niveau de risque est associé un quantile de la distribution du marqueur et vice versa. Cette interprétation dans les deux sens n'est possible que lorsque la relation entre le marqueur et le risque prédit est monotone croissante. Dans le cas contraire, une définition plus générale de la courbe de prédiction est nécessaire :

$$R(\nu) = p \ : \ P(\text{risque}(Y) \leq p) = \nu$$

Un trait horizontal est rajouté sur les graphiques : il correspond à la prévalence de la maladie. Un bon marqueur tend à ce que le risque prédit pour les futurs malades soit nettement supérieur à la prévalence et le risque prédit pour les futurs non malades soit nettement inférieur à la prévalence. C'est donc un marqueur dont la courbe de prédiction s'écarte fortement de la prévalence, en étant au dessus de la prévalence pour les risques élevés et en dessous de la prévalence pour les risques faibles.

Dans l'exemple de la figure 3.5, 84 % des patients ont un risque prédit en dessous de 90 % pour le marqueuré 1, ce qui signifie que 16 % ont un risque prédit au dessus de 90 %, alors que pour le marqueur 2, il n'y a que 4 % des patients qui ont un risque prédit au dessus de

90 % ; de même, 34 % de la population a un risque prédit en dessous de 1 % pour le marqueur 1, alors que pour le marqueur 2, uniquement 16 % des patients a un risque prédit en dessous de 1 %. Le marqueur 1 semble meilleur que le marqueur 2. La courbe de prédiction ne mesure pas réellement la capacité du test à discriminer les patients en deux groupes, mais à associer des risques extrêmes pour une grande partie de la population.

Une mesure d'adéquation souvent proposée pour les modèles de risque est le R carré, ou R^2, correspondant à la proportion de variation expliquée par le modèle (Mittlböck et Schemper, 1996). Celui-ci peut se calculer à partir de la courbe de prédiction, avec une interprétation très concrète :

$$R^2 = \int_0^1 (R(\nu) - \pi)^2 \, d\nu / (\pi(1-\pi))$$

π correspond à la prévalence de la maladie et $\pi(1-\pi)$ est un facteur de standardisation. Le R^2 est donc tout simplement le reflet de la capacité du marqueur à prédire des risques qui se démarquent fortement de la prévalence.

Il est possible de calculer les sensibilités, spécificités et valeurs prédictives pour chaque niveau du marqueur à partir de la courbe de prédiction (figure 3.6).

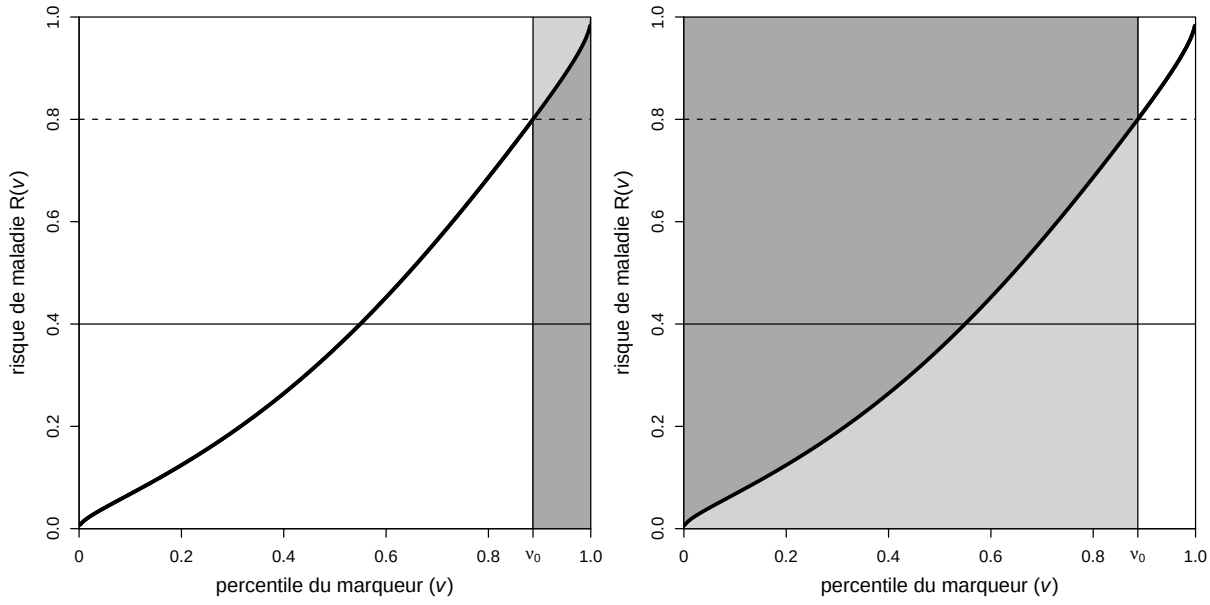

Figure 3.6 – Courbes de prédiction ainsi que leurs relations avec la sensibilité, la spécificité et les valeurs prédictives.

Pour un seuil correspondant au quantile ν_0 du marqueur, la valeur prédictive positive correspond à l'AROC représentée en gris foncé, divisée par l'aire du rectangle en gris clair

(graphique de gauche de la figure 3.6). En effet :

$$\mathrm{Vpp}(\nu_0) = P(M|Y > F^{-1}(\nu_0)) = \frac{\int_{\nu_0}^{1} P(M|Y = F^{-1}(\nu))\, d(P(Y \leq F^{-1}(\nu)))}{P(Y > F^{-1}(\nu_0))}$$

Le numérateur est donné par l'aire de la surface en gris foncé ; le dénominateur est quant à lui obtenu par l'aire du rectangle grisé. La sensibilité correspond à l'aire de la surface en gris foncé divisée par la prévalence. En effet, d'après le théorème de Bayes :

$$\mathrm{Sen}(\nu_0) = \frac{\mathrm{Vpp}(\nu_0) \times P(Y > F^{-1}(\nu_0))}{P(M)}$$

Les valeurs prédictives négatives et valeurs de spécificité se retrouvent de manière similaire : la valeur prédictive négative correspond à l'aire au dessus de la courbe représentée en gris foncé (graphique de droite de la figure 3.6) divisée par l'aire du rectangle grisé ; la spécificité est donnée par l'aire de la surface en gris foncé divisée par le complément de la prévalence. Ainsi, bien que le rôle fondamental de la courbe de prédiction ne soit pas l'étude de la capacité du marqueur à discriminer les patients, elle permet tout de même d'observer visuellement les mesures de performance pour la discrimination. Un marqueur avec de bonnes sensibilités est un marqueur pour lequel la courbe de prédiction monte rapidement vers le haut ; à l'inverse, un marqueur avec de bonnes spécificités a une courbe de prédiction qui reste le plus possible vers le bas.

En plus de l'observation de l'évolution du risque prédit en fonction du niveau du marqueur, les courbes de prédiction permettent d'analyser l'utilité du marqueur lorsqu'il est appliqué à une population, en calculant les pourcentages d'individus dont les risques prédits atteignent des niveaux faibles ou élevés. Il faut donc définir deux risques limites : un au dessus duquel les patients sont jugés comme à risque élevé et l'autre en dessous duquel les patients sont jugés à risque faible. Dans le cas de la figure 3.5, si ces risques limites sont respectivement de 10 % et 90 %, la conclusion, suite à la mesure du marqueur, est indécise pour 50 % des patients en utilisant le marqueur 1 (risque prédit entre les deux risques limites) et pour 86 % de la population pour le marqueur 2. Le marqueur 1 semble donc plus adapté dans ce cas, l'enjeu étant que le minimum de personnes pour lesquelles une décision tranchée n'est pas envisageable ait un risque prédit intermédiaire.

Le choix de ces valeurs limites de risque dépend du contexte clinique et revient à quantifier les coûts et les bénéfices associés au fait de classer les patients à risque élevé et risque faible.

Ainsi, encore une fois, il est impossible de comparer des marqueurs en se basant uniquement sur la distribution de probabilité conjointe des valeurs de marqueur et du statut des patients.

Les courbes de prédiction constituent un outil utile pour analyser les risques prédits par un marqueur et la capacité de celui-ci à discriminer les patients. Toutefois Margaret Pepe souligne que leur construction est plus complexe que celle de courbes ROC, les premières nécessitant d'estimer des densités de probabilité, les secondes ne nécessitant l'estimation que de fonctions de répartition (Pepe et *al.*, 2008).

3.4 Introduction de l'utilité

L'utilité espérée d'un marqueur dans une population est donnée par la formule (2.4). Elle est calculable à partir des factorisations orientées marqueur et patient. Néanmoins, les mesures de performance associées à ces deux types de factorisation n'intègrent pas directement la notion d'utilité, ce qui limite leur capacité à faire le choix entre plusieurs marqueurs.

Une difficulté liée à la notion d'utilité est qu'elle est relative à la personne concernée ; ainsi, il n'y a pas une seule utilité espérée pour un marqueur, mais un ensemble d'utilités dépendant des préférences des patients et des cliniciens. Des outils sont donc nécessaires pour sélectionner un marqueur non pas sur une seule mesure d'utilité, mais pour un ensemble de mesures d'utilité représentant les variations de préférences existant au sein de la population cible.

3.4.1 Courbes ROC et courbes de risque

Les courbes ROC n'incluent pas directement la notion d'utilité. Néanmoins, elles représentent les sensibilités et spécificités qui interviennent dans le calcul de l'utilité espérée, si la factorisation orientée marqueur est retenue (équation 2.5). Si les courbes ROC associées à deux marqueurs ne se croisent pas, alors :

- quel que soit le ratio bénéfice net sur coût net choisi,
- quelle que soit la prévalence,
- et quel que soit le seuil de positivité choisi,

l'utilité espérée à partir du marqueur ayant l'AROC la plus élevée est toujours supérieure à celle du second marqueur. Dans ce cas, il n'est pas nécessaire de déterminer le ratio bénéfice net sur coût net pour le choix du marqueur ; ce choix reste valable pour des populations aux prévalences différentes. Ce n'est plus le cas lorsque les courbes ROC se croisent ; selon les valeurs de spécificité, l'un des marqueurs est plus utile que l'autre et vice versa. Ainsi, l'utilité d'un des

marqueurs n'est pas toujours supérieure à celle des autres indépendamment de la prévalence et du ratio bénéfice net sur coût net. La spécification de ces deux derniers paramètres est nécessaire pour le choix du marqueur. Des AROC partielles ont été proposées (Pepe, 2003), consistant à ne comparer l'AROC que pour une zone de valeurs de spécificité qui présentent un intérêt pour le problème considéré. Parfois, ceci permet de restreindre les courbes ROC à des portions où elles sont emboîtées, mais définir une zone de valeurs de spécificité d'intérêt revient indirectement à fixer la prévalence et le ratio bénéfice net sur coût net.

Les mesures de performance liées à l'étude de la calibration n'ont pas quant à elles pour objectif de classer les patients ; elles sont donc totalement indépendantes de l'action réalisée en fonction de la valeur du marqueur. Ainsi, l'intégration de la notion d'utilité dans ces mesures est difficile. L'utilité intervient indirectement dans les courbes de prédiction, dans le choix des limites de risque basses et élevées, mais ces limites peuvent varier d'une équipe de soin à l'autre, en fonction du ratio bénéfice net sur coût net. Un outil est donc nécessaire pour déterminer les gammes de valeurs de ratio bénéfice net sur coût net dans lesquelles le marqueur est utile. Ce sont les courbes de décision.

3.4.2 Courbes de décision

Les courbes de décision, présentées entre autres dans les articles de Vikers et Elkin (Vickers et Elkin, 2006) et de Vickers et *al.* (Vickers et *al.*, 2008), représentent l'utilité des différents marqueurs en fonction du ratio bénéfice net sur coût net choisi, pour une valeur de prévalence fixée. Ainsi, il est facile de déterminer dans quelle gamme de ratio bénéfice net sur coût net un marqueur est utile. En réalité, la représentation n'est pas effectuée en fonction du ratio bénéfice net sur coût net, mais en fonction de la probabilité de survenue de la maladie à partir de laquelle un patient accepterait d'être traité, ou à partir de laquelle un clinicien conseillerait le traitement. Cette probabilité et le ratio bénéfice net sur coût net sont deux façons équivalentes de quantifier les préférences des patients ou des cliniciens.

Supposons qu'il existe un modèle permettant de prédire le risque de maladie d'un patient en fonction de sa valeur de marqueur : si le risque prédit est proche de 1, le patient va probablement accepter le traitement ; à l'inverse, si le risque prédit est proche de 0, le patient va probablement le refuser ; entre les deux, il existe une valeur de risque pour laquelle le patient est indécis. Cette probabilité, notée par la suite r, est la probabilité de maladie pour laquelle l'utilité espérée en traitant un patient est égale à l'utilité espérée en ne traitant pas un patient.

Ainsi :

$$r \times u(z_{TM}) + (1-r) \times u(z_{T\bar{M}}) = r \times u(z_{\bar{T}M}) + (1-r) \times u(z_{\bar{T}\bar{M}})$$

Après quelques transformations, il peut être montré que :

$$\frac{BN}{CN} = \frac{1-r}{r} \quad \Leftrightarrow \quad r = \frac{1}{1+BN/CN}$$

Ainsi, les préférences des cliniciens et des patients sont quantifiables à la fois en termes de ratio bénéfice net sur coût net et en termes de probabilité ou risque à partir duquel le traitement est proposé ou accepté. Par la suite, l'une ou l'autre des notions sera utilisée pour quantifier les préférences. Néanmoins, il est à noter que le ratio bénéfice net sur coût net s'interprète essentiellement pour un ensemble d'individus, alors que la probabilité de maladie à partir de laquelle le traitement est accepté peut plus facilement s'interpréter à l'échelle d'un individu.

Vickers et Elkin ne calculent pas l'utilité à partir des formules 2.4 ou 2.5, mais à partir d'une fonction qui est maximisée pour les mêmes valeurs de sensibilité et de spécificité :

$$\text{Bénéfice espéré de l'action} = \text{Sen} \times \pi - (1-\text{Spe}) \times (1-\pi) \times \frac{r}{1-r} \tag{3.3}$$

Si le bénéfice associé à l'action de traiter un patient malade par rapport à ne pas le traiter est fixé arbitrairement à 1, alors le coût net associé au fait de traiter à tort un patient non malade est donné par $r/(1-r)$; ainsi, le bénéfice espéré en traitant un patient est donné par la relation (3.3). Un marqueur qui ne classerait aucun patient comme malade (et avec lequel aucun patient ne serait traité) aurait un bénéfice espéré du traitement nul.

En représentant le bénéfice espéré du traitement associé à l'utilisation de chaque marqueur en fonction du risque à partir duquel un patient accepte d'être traité (r), il est possible de déterminer dans quelles zones chacun des marqueurs est plus utile que les autres.

Lorsque qu'un marqueur prend des valeurs continues, le bénéfice espéré du traitement pour la valeur r est calculé à partir des estimations de sensibilité et de spécificité associées au seuil du marqueur qui maximise l'utilité pour la valeur r en question.

Ces courbes de décisions permettent d'introduire dans le choix du marqueur l'utilité associée à chacune des situations possibles suite à la mesure du marqueur, ainsi que la variabilité de ces utilités d'un patient ou d'un clinicien à l'autre.

3.5 Bilan du chapitre 3

Cette partie avait pour but de présenter différents outils pour évaluer les performances de marqueurs et de déterminer le marqueur à retenir. Le choix du type d'outil dépend de la question posée : souhaite-t-on former des groupes de patients, ou mesurer la capacité du marqueur à refléter le vrai risque de ces patients ? En règle générale, la question oscille entre ces deux extrêmes ; il est conseillé de comparer les résultats obtenus avec les différents outils. Les éléments présentés au cours de cette partie devraient permettre de mieux comprendre les éventuelles différences de résultats.

Chapitre

4 Estimation et choix de marqueur

L'objectif de cette partie est de comparer les capacités du nadir de PSA, de la date du nadir et de la vélocité à discriminer les patients en patients malades (avec au moins une biopsie positive) et patients non malades (aucune biopsie positive). Afin de s'affranchir des problèmes de fréquence des mesures et d'augmentations ponctuelles de PSA non liées à la persistance de cellules cancéreuses, les différents marqueurs ont été calculés à partir de la modélisation robuste des profils de PSA de chacun des patients.

Une fois les marqueurs estimés pour chaque patient, leurs performances ont été quantifiées au travers de courbes ROC. Ici, l'objectif est clairement d'identifier deux groupes de patients : l'un pour lesquels une biopsie est fortement recommandée et l'autre pour lesquels une biopsie n'est pas nécessaire. Ainsi, les mesures de discrimination ont été préférées aux mesures de calibration.

Ce chapitre décrit les méthodes qui ont été utilisées afin d'obtenir les estimations des marqueurs les moins biaisées possibles pour chacun des patients, ainsi que les résultats concernant la comparaison des performances des trois marqueurs considérés. Une dernière partie a été ajoutée afin de justifier de l'intérêt de la modélisation des profils de PSA des patients.

4.1 Méthodes

4.1.1 Choix d'un modèle pour décrire les profils de PSA des patients

Les profils de PSA des patients sont caractérisés par une forte diminution de PSA juste après le traitement, puis par une phase de stabilisation ou de ré-augmentation (figure 1.2). Par la suite, on appellera *trajectoire* de PSA l'évolution des valeurs de PSA qui seraient obtenues pour un patient si les mesures étaient journalières et non sujettes aux phénomènes d'augmentations ponctuelles non liées à la persistance de cancer.

Un modèle de décroissance exponentielle/croissance exponentielle a été choisi pour décrire les trajectoires, faisant intervenir quatre paramètres par patients :

$$y(t) = \exp(r_1)\exp(-r_2 t) + \exp(r_3 t)\exp(r_4 t) \tag{4.1}$$

$y(t)$ correspond à la mesure de PSA effectuée t jours après le traitement UFHI ; $\exp(r_1) + \exp(r_3)$ correspond à l'intercept de la trajectoire de PSA après le traitement, r_2 est la vitesse de décroissance des PSA et r_4 la vitesse de ré-augmentation. Bien que ce modèle ait été obtenu de manière empirique, il a, en partie, une interprétation biologique : la phase de croissance exponentielle des PSA ($\exp(r_4 t)$) reflète le fait que, en cas de persistance de cancer, la tumeur est supposée avoir une croissance exponentielle ; les taux de PSA sont donc proportionnels au volume de la tumeur (Taylor et *al.*, 2005).

D'autres modèles ont été proposés pour décrire la cinétique des PSA après un traitement de type radiothérapie ou ultrasons. You et *al.* (2008) ont, entre autres, décrit un modèle pharmacocinétique à deux compartiments, permettant de refléter deux processus d'élimination des PSA après le traitement : un premier très rapide et un second plus lent. Ces deux phases étaient difficilement observables sur les données de PSA après traitement par ultrasons ; de plus, ce modèle fait intervenir six paramètres par patient, alors que l'on ne disposait pas, en général, de suffisamment de mesures par patient pour estimer six paramètres.

Les modèles log-linéaires par morceaux avec un temps de changement de pente estimé à partir des données ont également souvent été utilisés pour décrire des profils de PSA après un traitement (Bellera et *al.*, 2008). Néanmoins, dans ce type de modèle, la jonction entre les phases de décroissance et de croissance des PSA est décrite par un brusque changement de pente, alors que pour les données de PSA après UFHI, cette jonction était clairement courbée. La zone de transition entre les deux phases d'évolution étant primordiale pour la détermination du nadir de

PSA, le modèle de décroissance exponentielle/croissance exponentielle, plus souple, a été préféré. Ce modèle a fourni un ajustement raisonnable des données pour la plupart des patients.

A partir du modèle (4.1) les trois marqueurs étudiés sont calculables théoriquement pour chaque patient grâce aux valeurs $\{r_1, r_2, r_3, r_4\}$ estimées par patient :

- pour la date du nadir : $T = \frac{1}{r_2+r_4} \ln\left(\frac{\exp(r_1)r_2}{\exp(r_3)r_4}\right)$;
- pour le nadir : $N = \exp(r_1)\exp(-r_2 T) + \exp(r_3)\exp(r_4 T)$
- enfin la vélocité est directement obtenue à partir du paramètre r_4.

Le modèle (4.1) est le modèle théorique des valeurs de PSA. Dans la réalité, il existe toujours un écart entre la valeur prédite et la valeur observée ; si y_{ij} dénote la $j^{\text{ième}}$ mesure du patient i effectuée au temps t_{ij} (avec m_i mesures pour le patient i), le modèle aléatoire est donné par :

$$y_{ij} = \psi(t_{ij}, \boldsymbol{r}_i) + \varepsilon_{ij} \tag{4.2}$$

où $\boldsymbol{r}_i = (r_{1i}, r_{2i}, r_{3i}, r_{4i})^T$ est le vecteur des paramètres pour le patient i et $\psi(t_{ij}, \boldsymbol{r}_i)$ est la valeur de PSA prédite par le modèle (4.1) pour le patient i au temps t_{ij}. Les résidus ε_{ij} sont supposés suivre une distribution gaussienne.

En raison du faible nombre de mesures par patient, le modèle n'a pas pu être ajusté séparément pour chacun d'eux. Un modèle à effets mixtes a donc été utilisé pour estimer les paramètres de chaque patient. Celui-ci fait l'hypothèse que les effets aléatoires suivent une distribution normale multivariée sur l'ensemble des individus. En raison, encore une fois, du nombre limité de mesures, la corrélation entre les différents effets aléatoires n'a pas pu être modélisée. Le modèle général était donc de la forme :

$$\begin{gathered} y_{ij} = \exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \hookrightarrow \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \\ r_{1i} \hookrightarrow \mathcal{N}\left(\mu_{r_1}, \sigma_{r_1}^2\right), \quad r_{2i} \hookrightarrow \mathcal{N}\left(\mu_{r_2}, \sigma_{r_2}^2\right), \quad r_{3i} \hookrightarrow \mathcal{N}\left(\mu_{r_3}, \sigma_{r_3}^2\right), \quad r_{4i} \hookrightarrow \mathcal{N}\left(\mu_{r_4}, \sigma_{r_4}^2\right) \end{gathered} \tag{4.3}$$

Une analyse de modèles emboîtés a montré que l'utilisation d'un effet aléatoire pour chacun des paramètres du modèle, par rapport à un simple effet fixe, améliorait significativement la vraisemblance (tableau 4.1). Le modèle a effets mixtes ne convergeait que pour les patients ayant au moins cinq mesures de PSA, c'est pourquoi, pour des raisons calculatoires, les patients ayant moins de cinq mesures de PSA ont été retirés de l'analyse.

Le fait que la corrélation entre les effets aléatoires n'ait pas pu être modélisée pourrait être dû à une sur-paramétrisation du modèle et non à un manque de données par patient. Néanmoins,

Tableau 4.1 – Comparaison des log-vraisemblances des modèles en incluant plus ou moins d'effets aléatoires.

Effets fixes	Effets aléatoires	Log-vraisemblance	valeur P pour le ratio de vraisemblance
r_1, r_2, r_3, r_4		-2285	
r_2, r_3, r_4	r_1	-2023	$< 0,0001$
r_3, r_4	r_1, r_2	-31	$< 0,0001$
r_4	r_1, r_2, r_3	-290	$< 0,0001$
	r_1, r_2, r_3, r_4	-28	$< 0,0001$

en simulant des données de PSA pour 289 patients hypothétiques à partir du modèle (4.3), les corrélations entre les effets aléatoires étaient estimables lorsque huit mesures étaient générées par patient. Ceci n'était plus le cas lorsque le nombre de mesures par patient était inférieur à sept, le nombre médian de mesures de PSA par patient dans les vrais données de PSA étant de sept. Ainsi, il a été considéré que le manque de mesures par patient était à l'origine de l'impossibilité de modéliser la corrélation entre les effets aléatoires.

4.1.2 Variabilité intra-patients

4.1.2.1 Distribution des PSA

La distribution des valeurs de PSA était asymétrique, avec une asymétrie vers la droite. Selon Carroll et Ruppert (1988), des données asymétriques peuvent être modélisées en transformant les deux côtés de l'équation (4.1) à l'aide d'une même fonction. Dans le cadre des données de PSA, la famille de transformations de Box-Cox a été considérée. Cette famille fait intervenir un paramètre λ :

$$u_{ij} = \begin{cases} (y_{ij}^{\lambda} - 1)/\lambda & \text{si } \lambda \neq 0 \\ \ln(y_{ij}) & \text{si } \lambda = 0 \end{cases}$$

Les données transformées u_{ij} suivent ainsi une distribution gaussienne. Une adaptation de la méthode de Lipsitz et *al.* (2000) a été utilisée pour estimer la valeur du paramètre λ dans le cadre de données longitudinales.

La densité de probabilité des observations transformées est donnée par :

$$P(\mathbf{u}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2) = \int P(\mathbf{u}_i|\mathbf{r}_i, \sigma_\varepsilon^2) P(\mathbf{r}_i|\boldsymbol{\Sigma}, \boldsymbol{\mu}, \sigma_\varepsilon^2)\, d\mathbf{r}_i \tag{4.4}$$

où $\boldsymbol{\mu} = (\mu_{r_1}, \mu_{r_2}, \mu_{r_3}, \mu_{r_4})^T$ et $\boldsymbol{\Sigma}$ est la matrice diagonale de variance covariance des effets aléatoires. Les deux termes sous l'intégrale sont des densités de probabilité de lois normales. La densité de probabilité des vraies observations $\mathbf{y}_i$ est donnée par la densité de probabilité (4.4) multipliée par :

$$J_i = \prod_{j=1}^{n_i} |y_{ij}^{\lambda-1}|$$

J_i correspondant à la valeur absolue de la jacobienne de la transformation qui à y_{ij} associe u_{ij}. Ainsi, la log-vraisemblance se décompose en deux termes :

$$\ln(L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \lambda)) = \ln(L_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \lambda)) + \ln(L_2(\lambda)) \tag{4.5}$$

avec

$$\ln(L_1(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2, \lambda)) = \sum_{i=1}^{n} \ln(P(\mathbf{u}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \sigma_\varepsilon^2), \lambda)$$

et

$$\ln(L_2(\lambda)) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \ln |y_{ij}^{\lambda-1}|$$

L'estimation de λ est effectuée par vraisemblance profilée, en calculant (4.5) pour différentes valeurs du paramètre λ (figure 4.1).

La valeur optimale de λ était comprise entre -0,5 et 0 et a été arrondie à 0. Ceci revient à considérer une transformation logarithmique, très souvent utilisée dans la littérature sur les PSA, mais rarement justifiée. L'erreur de mesure est ainsi multiplicative et non additive. Une valeur de 1 a été ajoutée dans la transformation logarithmique pour éviter les problèmes de valeurs nulles de PSA. Le modèle ainsi obtenu est donné par :

$$\ln(y_{ij} + 1) = \ln(\exp(r_{1i}) \exp(-r_{2i} t_{ij}) + \exp(r_{3i}) \exp(r_{4i} t_{ij}) + 1) + \varepsilon_{ij}, \quad \varepsilon_{ij} \hookrightarrow \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$$

4.1.2.2 Prise en compte des valeurs aberrantes

Le graphique des résidus standardisés en fonction des valeurs prédites figurant dans l'article publié dans Statistics in Medicine met clairement en évidence un certain nombre de valeurs aberrantes. Celles-ci correspondent majoritairement aux fortes augmentations de PSA de très courtes durées, indépendantes du cancer, mais liées, par exemple, à de simples infections ou inflammations. L'amplitude de ces variations les rend difficilement compatibles avec une variabilité de méthodes de mesures entre les laboratoires d'analyses. Ces augmentations n'étant pas en lien

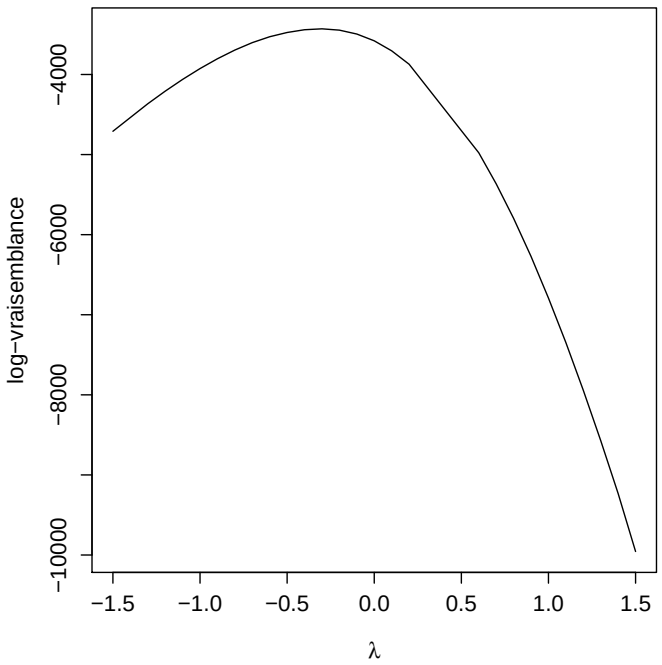

Figure 4.1 – Log-vraisemblance des données de PSA selon le modèle (4.3) en fonction de la valeur de λ.

avec la persistance de cellules cancéreuses, il a été choisi de ne pas les modéliser. Néanmoins, il a fallu en tenir compte, car l'inférence sur les paramètres d'un modèle, lorsqu'une loi normale est supposée pour les résidus, peut être biaisée par la présence de valeurs aberrantes (Lange et *al.*, 1989 ; Pinheiro et *al.*, 2001).

La distribution de Gauss pour les résidus a par conséquent été remplacée par une distribution de Student, qui inclut un paramètre de plus que la loi normale : ν, le nombre de degrés de liberté. La loi de Student a des queues de distribution beaucoup plus larges que celles de la loi normale ; elle peut donc tolérer la présence de valeurs aberrantes. Plus le nombre de degrés de liberté est élevé, plus la loi de Student se rapproche d'une loi normale. Ce nombre de degrés de liberté est estimable à partir des données ; il représente l'importance des valeurs aberrantes dans les données. L'utilisation de la loi de Student revient à pondérer la participation des données à la vraisemblance avec un poids inversement proportionnel au carré de l'écart entre la valeur observée et la valeur prédite (Lange et *al.*, 1989). Les valeurs aberrantes, pour lesquelles cette distance est élevée, participent donc moins à l'estimation des paramètres des patients. Un exemple est présenté dans l'article, où la trajectoire prédite en utilisant une loi de Gauss est fortement attirée par une valeur de PSA anormalement élevée, alors qu'elle l'est beaucoup moins en utilisant une loi de Student.

Le modèle ainsi obtenu est donné par :

$$\ln(y_{ij}+1) = \ln(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1) + \varepsilon_{ij}, \quad \varepsilon_{ij} \hookrightarrow t\left(\sigma_\varepsilon^2, \nu\right) \quad (4.6)$$

4.1.3 Inférence Bayésienne

Les paramètres de tous les modèles construits ont été estimés en utilisant l'inférence Bayésienne. Elle n'apporte rien de plus que l'inférence fréquentiste pour les modèles décrits précédemment, mais sera indispensable dans la partie concernant l'estimation du seuil optimal du marqueur retenu. Une introduction succincte de l'inférence Bayésienne est proposée en annexe (partie A). Une présentation plus détaillée et complète peut être trouvée dans Gilks et *al.* (1996) et Gelman et *al.* (2004).

4.1.4 Variabilité inter-patients

La partie 4.1.2 a permis d'assouplir l'hypothèse de normalité des résidus du modèle (4.3). Une autre hypothèse de ce modèle est que la distribution des effets aléatoires sur l'ensemble des patients est gaussienne. D'après les graphiques quantiles quantiles associés aux estimations des effets aléatoires du modèle (4.3), cette hypothèse s'est révélée être non valide, au moins pour les effets aléatoires r_1 et r_4 (graphique figurant dans l'article publié dans Statistics in Medicine). Ceci n'est pas totalement aberrant, car la population est constituée d'au moins deux sous-populations : celle avec et celle sans persistance de cellules cancéreuses. Ce mélange de populations peut se refléter au niveau des effets aléatoires par un mélange de deux distributions, ou plus. L'hypothèse de normalité des effets aléatoires a donc été assouplie afin de ne pas contraindre les estimations des effets aléatoires, conduisant ainsi à une estimation plus fiable des marqueurs. Des méthodes de modélisation non paramétriques ont été employées. Elles consistent à caractériser la distribution par un ensemble de points ayant chacun une certaine probabilité (mass points en anglais), en donnant une information a priori sur le nombre de points et sur la distribution plausible. Ceci correspond aux processus de Dirichlet.

4.1.4.1 Processus de Dirichlet

Les processus de Dirichlet sont couramment utilisés en Bayésien pour assouplir les hypothèses concernant la modélisation de la distribution d'une variable aléatoire. Une présentation détaillée peut être trouvée dans Escobar (1994) et Escobar et West (1995) ; une présentation plus succincte et moins mathématique, mais plus intuitive, est proposée ci-après.

Lorsqu'une variable aléatoire ne peut prendre que deux valeurs, 0 ou 1, et que cette variable aléatoire est mesurée n fois, le nombre de fois où la valeur 1 est obtenue (y) peut être modélisé à l'aide d'une loi binomiale de paramètre n et θ, θ correspondant à la probabilité d'obtenir la valeur 1. Classiquement, une loi beta de paramètres a et b est utilisée pour donner de l'information a priori sur θ. Dans ce cas, la distribution a posteriori de θ est une loi beta de paramètres $a + y$ et $b + n - y$. L'information a priori rajoute, par rapport aux données, a événements sur un ensemble de $a + b$ mesures.

On considère maintenant le cas où la variable aléatoire peut prendre plus de deux valeurs. Par exemple, au cours d'un lancer de dés, six valeurs différentes peuvent être obtenues. Soit $\theta_i, i = 1, \ldots, 6$ la probabilité d'obtenir chacune de ces valeurs. La généralisation de la loi beta aux cas de plus de deux événements possibles est une loi de Dirichlet, de paramètres M et G_0. G_0 est l'espérance des valeurs de θ ; a priori, si le dé n'est pas pipé, on peut supposer que $G_0 = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)^T$. M est un paramètre de précision autour de G_0 ; plus la valeur de M est élevée, plus G_0 est supposé fiable pour les valeurs de θ. A l'issue de n tirages de dé, la distribution des paramètres θ est également une distribution de Dirichlet, de paramètres $M^\star = M + n$ et $G = (MG_0 + nF)/(M + n)$, où F correspond aux proportions de chacune des six valeurs obtenues durant les n tirages.

Considérons maintenant une variable aléatoire Y prenant des valeurs continues. L'espace des valeurs possibles pour Y peut être subdivisé en une infinité de sous-espaces $X_j, j = 1, \ldots, k$ (figure 4.2). Soit θ_j la probabilité que Y appartienne au petit intervalle X_j. Les valeurs de θ pour l'ensemble des intervalles :

- peuvent être modélisées à l'aide d'une distribution de Dirichlet ;
- forment une distribution de probabilité pour Y, notée G, lorsque k tend vers l'infini.

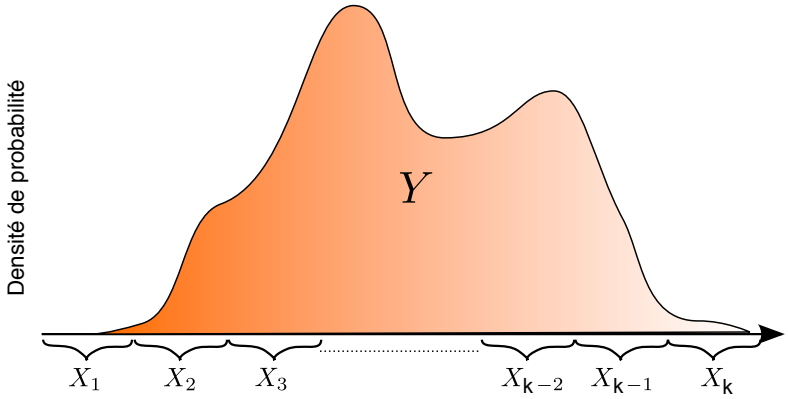

Figure 4.2 – Densité de probabilité de la variable aléatoire Y, l'espace des valeurs possibles étant subdivisé en une infinité de sous-espaces $X_j, j = 1, \ldots, k$.

Ainsi :

$$\begin{aligned} Y|G &\hookrightarrow G \\ G &\hookrightarrow \mathcal{DP}(M, G_0) \end{aligned}$$

où $\mathcal{DP}$ dénote un processus de Dirichlet, G est la distribution de Y, G_0 est la distribution supposée a priori de Y et M est un paramètre caractérisant la confiance en G_0.

Les processus de Dirichlet sont une simple extension de la loi de Dirichlet au cas de données continues. Ils sont considérés comme des équivalents Bayésiens des méthodes fréquentistes d'estimation non paramétrique de densité de probabilité de type NPMLE (Laird, 1978). Ils ont déjà été utilisés par un certain nombre d'auteurs pour assouplir la distribution d'effets aléatoires dans le cas de modèles à effets mixtes, conduisant ainsi à des modèles mixtes semi-paramétriques (Bush et MacEachern, 1996 ; Kleinman et Ibrahim, 1998 ; Brown et Ibrahim, 2003 ; Ohlssen et *al.*, 2007). Les processus de Dirichlet peuvent être vus comme une façon de donner un a priori non pas sur les valeurs probables d'un paramètre, mais sur la distribution d'un ensemble de paramètres.

L'utilisation des processus de Dirichlet dans une chaîne MCMC peut s'effectuer par la technique dite des " restaurants chinois ". L'implémentation algorithmique des processus de Dirichlet donne une autre vision de ceux-ci, c'est pourquoi elle est présentée de manière succincte ci-après.

Considérons un modèle faisant intervenir un effet aléatoire r. Soit $\mathbf{y}_i$ les mesures associées à un patient i et r_i la valeur de l'effet aléatoire pour ce patient i. L'effet aléatoire suit sur l'ensemble des patients une distribution G sur laquelle est placée une information a priori sous la forme d'un processus de Dirichlet de paramètres M et G_0. L'analogie de l'algorithme d'échantillonnage avec les restaurants chinois est décrite ci-après. Lorsqu'une personne seule rentre dans un bar, elle peut décider soit de s'asseoir à une table où des personnes sont déjà assises, en privilégiant les tables avec le plus de personnes pour ne pas rester seule – c'est l'effet clustering – soit choisir une table vide. Le même principe est utilisé pour échantillonner la valeur de l'effet aléatoire pour un patient i sachant les valeurs attribuées aux autres patients.

Le patient en question peut se voir attribuer une valeur d'effet aléatoire déjà attribuée à d'autres personnes, avec une probabilité égale à $n_k \times q_k$, où n_k est le nombre de patients appartenant au cluster k et $q_k = P(\mathbf{y}_i|r_k)$. Il peut également se voir attribuer une nouvelle valeur d'effet aléatoire, issue de G_i, où G_i est la distribution a posteriori de $r_i|\mathbf{y}_i$ selon un a

priori G_0 (i.e. $G_i(r_i) \propto P(\mathbf{y}_i|r_i)G_0(r_i)$), et ce, avec une probabilité q_0 donnée par :

$$q_0 = M \int P(\mathbf{y}_i|r_i)G_0(r_i)\, dr_i$$

Ainsi :

$$P(r_i|r_{-i}, \mathbf{y}_i) \propto \sum_{\substack{k=1 \\ k\neq i}}^{n} q_k \delta(r_k) + q_0 G_i(r_i)$$

$\delta(r_k)$ est une fonction binaire, valant 1 quand $r_i = r_k$ et 0 partout ailleurs ; r_{-i} correspond aux valeurs d'effets aléatoires pour l'ensemble des patients à l'exception du patient i. Ainsi, r_i prend soit la valeur qui est la plus vraisemblable d'après $\mathbf{y}_i$ parmi les valeurs de r affectées aux autres patients, soit la valeur issue de G_0 la plus vraisemblable d'après $\mathbf{y}_i$. $P(r_i|r_{-i}, \mathbf{y}_i)$ est donc un mélange entre une distribution continue et des distributions discrètes. Plus M est élevé, plus la distribution des effets aléatoires se rapproche de la distribution a priori. En fonction de la vraisemblance et de la distribution choisie pour G_0, le calcul de q_0 peut ne pas être réalisable explicitement. Des solutions ont été proposées dans ce cas par Neal (2000). Enfin le paramètre M n'est pas forcément fixé ; il peut être estimé à partir des données en spécifiant une information a priori sur sa distribution. Une solution utilisant une loi gamma comme a priori est proposée par Escobar et West (1995).

4.1.4.2 Modèle Student/Dirichlet

Des processus de Dirichlet ont été utilisés pour modéliser la distribution des effets aléatoires r_1 et r_4. Par souci de parcimonie, la distribution des effets aléatoires r_2 et r_3 a été caractérisée par des lois normales, ces deux effets aléatoires semblant avoir des distributions gaussiennes. Le modèle ainsi obtenu est donné par :

$$\begin{gathered}
\ln(y_{ij}+1) = \ln\left(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1\right) + \varepsilon_{ij}, \quad \varepsilon_{ij} \hookrightarrow t\left(\sigma_\varepsilon^2, \nu\right) \\
r_{1i} \hookrightarrow \mathcal{DP}\left(M_1, G_{01}\right), \quad r_{2i} \hookrightarrow \mathcal{N}\left(\mu_{r_2}, \sigma_{r_2}^2\right), \quad r_{3i} \hookrightarrow \mathcal{N}\left(\mu_{r_3}, \sigma_{r_3}^2\right), \quad r_{4i} \hookrightarrow \mathcal{DP}\left(M_4, G_{04}\right) \\
G_{01} \hookrightarrow \mathcal{N}\left(\mu_{r_1}, \sigma_{r_1}^2\right), \quad G_{04} \hookrightarrow \mathcal{N}\left(\mu_{r_4}, \sigma_{r_4}^2\right)
\end{gathered} \tag{4.7}$$

Deux distributions normales ont été choisies comme espérance a priori des distributions des effets aléatoires r_1 et r_4 (G_{01} et G_{04}) ; des a priori non informatifs ont été utilisés pour l'ensemble des paramètres du modèle. Ce dernier modèle a été appelé modèle “ Student/Dirichlet ” (Student pour les résidus, processus de Dirichlet pour les effets aléatoires), par opposition au modèle (4.3) qualifié de “ Gauss/Gauss ”.

L'ensemble des paramètres a été échantillonné grâce à l'échantillonneur de Gibbs (annexe A). Les résultats de 3000 itérations ont été retenus. Chaque itération a donné lieu à une estimation d'effets aléatoires par patient ($(r_{1i}, r_{2i}, r_{3i}, r_{4i})^T$), donc à une valeur de chacun des marqueurs à partir des formules données dans la partie 4.1.1. Ainsi, 3000 courbes ROC ont été construites pour chacun des marqueurs ; la moyenne de la distribution a posteriori de l'AROC ainsi obtenue a été retenue comme estimation de l'AROC pour chaque marqueur.

4.1.5 Comparaison des modèles Student/Dirichlet et Gauss/Gauss

Le modèle Student/Dirichlet est censé assouplir les hypothèses de normalité des effets aléatoires et des résidus par rapport au modèle Gauss/Gauss, afin de conduire à une estimation moins biaisée des effets aléatoires par patient. Un ensemble de simulations a été réalisé afin de comparer les résultats obtenus à partir des deux modèles en présence de valeurs aberrantes dans les mesures et d'effets aléatoires non gaussiens, et ce, en termes de biais, de précision et de probabilité de couverture de l'intervalle de crédibilité à 95 % des effets aléatoires. Les données ont été simulées à partir d'un modèle un peu plus simple que le modèle des PSA, ne faisant intervenir que deux effets aléatoires.

Plusieurs scénarios ont été envisagés, en faisant varier l'amplitude et la proportion des valeurs aberrantes, ainsi que le degré de non-normalité des effets aléatoires. Ces scénarios sont décrits précisément dans l'article publié dans Statistics in Medicine.

De plus, si l'effet aléatoire est utilisé comme marqueur diagnostique d'une maladie, les AROC obtenues à partir des estimations provenant du modèle Student/Dirichlet et du modèle Gauss/Gauss ont été comparées en termes de biais et de précision.

4.2 Article publié dans Statistics in Medicine

Les résultats de l'article publié dans Statistics in Medecine ne sont basés que sur 285 patients, dont 146 avec au moins une biopsie positive, alors que l'étude comporte 289 patients. Au moment de la rédaction de l'article, les données de quatre patients étaient incomplètes et n'ont pas pu être incluses dans l'analyse.

L'annexe en ligne de l'article décrivant les distributions conditionnelles des paramètres est fournie en annexe B.

4.2.1 Article

Research Article

Statistics in Medicine

Received 13 January 2009, Accepted 5 November 2009 Published online 4 January 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3816

Robust non-linear mixed modelling of longitudinal PSA levels after prostate cancer treatment

F. Subtil[a,b,c,d]*† and M. Rabilloud[a,b,c,d]

The objective of this study was to develop a robust non-linear mixed model for prostate-specific antigen (PSA) measurements after a high-intensity focused ultrasound (HIFU) treatment for prostate cancer. The characteristics of these data are the presence of outlying values and non-normal random effects. A numerical study proved that parameter estimates can be biased if these characteristics are not taken into account. The intra-patient variability was described by a Student-*t* distribution and Dirichlet process priors were assumed for non-normal random effects; a process that limited the bias and provided more efficient parameter estimates than a classical mixed model with normal residuals and random effects. It was applied to the determination of the best dynamic PSA criterion for the diagnosis of prostate cancer recurrence, but could be used in studies that rely on PSA data to improve prognosis or compare treatment efficiencies and also with other longitudinal biomarkers that, such as PSA, present outlying values and non-normal random effects. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: prostate-specific antigen; longitudinal study; robust modelling; semi-parametric Bayesian methods; diagnostic test

1. Introduction

The prostate-specific antigen (PSA) is a widely used biomarker for patient follow-up after prostate cancer. Its level is assumed to drop quickly after a given treatment until reaching a nadir. Generally, a significant rise in the PSA level after treatment leads to a more active monitoring of patients. Some PSA-based endpoints have been defined as surrogate for overall survival [1] or prostate cancer-specific mortality [2]. In clinical current practice, after radiotherapy, the Phoenix criteria [3], based in part on PSA levels, are commonly used to make the diagnosis of treatment failure or cancer recurrence. After high-intensity focused ultrasound (HIFU) treatment, there is still no agreement on the criteria that should be used for diagnosing recurrence even if some PSA endpoints have been suggested [4, 5]. Clinicians need such criteria to initiate biopsies and, depending on its results, adapt the therapy. From a cohort of patients followed at Hôpital Edouard Herriot (Lyon, France) after an HIFU treatment [6], the present study compares the diagnostic accuracy of some dynamic PSA criteria: the nadir, the time to this nadir, or the rate of PSA increase after the first decrease, called PSA velocity.

These criteria may be directly calculated from longitudinal observed data but then their values may depend on measurement errors, temporal fluctuations, and number of measurements. Typically, the estimates of the nadir and the time to the nadir depend strongly on the measurement frequency. Thus, it seems more relevant to model the longitudinal measurements obtained from each patient, and then estimate the dynamic criteria from the model so as to limit the influence of the aforementioned sources of variability. One contribution of this article is to propose a relevant model for longitudinal PSA measurements after the treatment.

PSA measurements were modelled using a non-linear mixed model, with random effects on the parameters of the model. One feature of PSA measurements, in this and other studies, is their occasional fluctuations owing to factors such as infection or prostatic manipulation [7, 8]. These important deviations from the expected PSA trend led to outliers in the model residuals.

[a] *Université de Lyon, F-69000 Lyon, France*
[b] *Université Lyon 1, F-69100 Villeurbanne, France*
[c] *CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France*
[d] *Hospices Civils de Lyon, Service de Biostatistique, F-69003 Lyon, France*
* *Correspondence to: F. Subtil, Hospices Civils de Lyon, Service de Biostatistique, 162 Avenue Lacassagne, F-69003 Lyon, France.*
† *E-mail: fabien.subtil@chu-lyon.fr*

Contract/grant sponsor: French 'Ligue contre le Cancer'

Statistics
in Medicine

Moreover, because there are at least two populations—patients with recurrence (diseased) and patients without recurrence (non-diseased)—some of the random effect distributions might result from mixtures of distributions, contradicting thus the classical hypothesis of normal distribution of random effects. These two remarks may lead to a bias in random effect estimates, and hence to a poor modelling of the individual longitudinal measurements. In the present study, an accurate estimate of the individual parameters was needed to assess their accuracy as diagnostic test of recurrence. Instead of a Gaussian, a Student-*t* distribution was assumed for residuals to reduce the weight of the outliers in parameter estimates and a Dirichlet process was used to relax the hypothesis of normal random effects. This led to a Student (residuals)/Dirichlet (random effects) model instead of a Gauss/Gauss one. The parameters of both models were estimated using a Gibbs sampler algorithm [9].

This article shows how to use these methods to reduce bias in random effect estimates in case of outlying values and non-normal random effects. A comparison between the performance characteristics of the Gauss/Gauss and the Student/Dirichlet models is carried out under different patterns of residuals and random effects. This work was applied to PSA longitudinal measurements after an HIFU treatment.

2. The non-linear mixed model

2.1. Study description

The specific study involved patients who were offered salvage HIFU between 2000 and 2007 as a definitive local therapy for prostate cancer (Hôpital Edouard Herriot, Lyon, France). Patients' follow-up after therapy included PSA measurements every 3 months during the first year and every 6 months thereafter. Depending on the clinicians' usual practices, biopsies were carried out in case of PSA rise. We retained only 285 patients whose first treatment was HIFU and who had at least one biopsy 3 months after treatment; this corresponded to patients with suspected recurrence. These patients were aged 50–81 years (median 71); most had a localized prostate cancer (stage T1–T2) except nine with T3 stage. Among all patients kept, 146 who had at least one positive biopsy during follow-up were considered as having a cancer recurrence (diseased patients); the others were considered recurrence free (non-diseased patients). Biopsy results were used as gold standard to diagnose a recurrence of prostate cancer. The mean and median follow-up durations were 1112 and 958 days in non-diseased patients; the mean and median time to first positive biopsy in diseased ones were 505 and 269 days. In diseased patients, all PSA measurements made before the first positive biopsy were kept for analysis. In non-diseased patients only those made before the last negative biopsy were kept.

2.2. PSA trajectories

Figure 1 is a plot of post-treatment PSA measurements for 30 randomly selected patients. After therapy, there was first a clear pattern of PSA quick decline, corresponding to the treatment effect, sometimes followed by a slow increase. Besides this general pattern, there were some occasional fluctuations (solid lines in Figure 1) possibly due to inherent biological PSA variability or to factors such as infection or prostate manipulation or measurement errors [7, 8] but not to progression towards recurrence. Thus, we restricted ourselves to modelling the general PSA pattern of each patient, later called 'individual PSA trajectory'; the fluctuations were considered as a part of the residuals of the model.

Individual PSA trajectories were modelled using a non-linear exponential decay–exponential growth model [10, 11]:

$$y(t) = \exp(r_1)\exp(-r_2 t) + \exp(r_3)\exp(r_4 t) \tag{1}$$

In this formula, t denotes the time elapsed since treatment and $y(t)$ the PSA value at time t; $\exp(r_1)+\exp(r_3)$ corresponds to the intercept of the post-treatment PSA trajectory, r_2 is the rate of the PSA decrease after treatment, and r_4 the rate of the PSA increase after decrease. Although this model was empirically derived, it allows some interesting biological interpretations; in particular, the exponential growth part, $\exp(r_4 t)$, reflects the fact that, in prostate cancer recurrence, the tumours are supposed to have an exponential growth and the PSA level is thought to be proportional to tumour volume.

The same model has been previously used for PSA measurements after radiotherapy [12, 13]. Other authors have already described the PSA decline using a two-compartment pharmacokinetic model suggesting two processes of PSA elimination: a very fast and a slightly slower one [14]. However, in our study, there were too few measurements to estimate the parameters per patient of the previous pharmacokinetic model and no clear two-process PSA elimination. Moreover, a model with four parameters fitted well the data from most patients (see Figure 2 for four randomly chosen patients); thus, it seemed unnecessary to introduce additional parameters in this particular study.

Changepoint models [15] are another kind of models describing PSA measurements after treatment. They assume that longitudinal PSA measurements can be described, on the logarithmic scale, by a succession of segments; more precisely, one for the decay part, followed by another, with a different slope, for the growth part. However, this model was too rigid for our data, especially for the link between the decay and the growth parts; it supposed a steep slope change whereas, in our data, the link between the decay and the growth parts was curved; moreover, the growth part was not linear (on the log scale), but curved. We have consequently preferred the aforementioned exponential model.

2.3. The non-linear mixed model

The previous model may be separately fitted for each patient. However, half of them had no more than seven measurements, leading to poor accuracy estimates of the four parameters. Therefore, all PSA measurements were included in a mixed-model [16], considering each parameter $\{r_1, r_2, r_3, r_4\}$ as a random effect to reflect inter-patients variability.

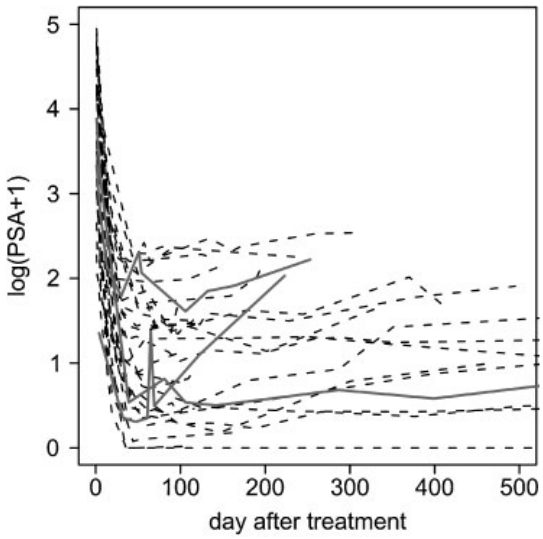

Figure 1. Plot of the logarithm of PSA values according to the day after treatment for 30 randomly selected patients.

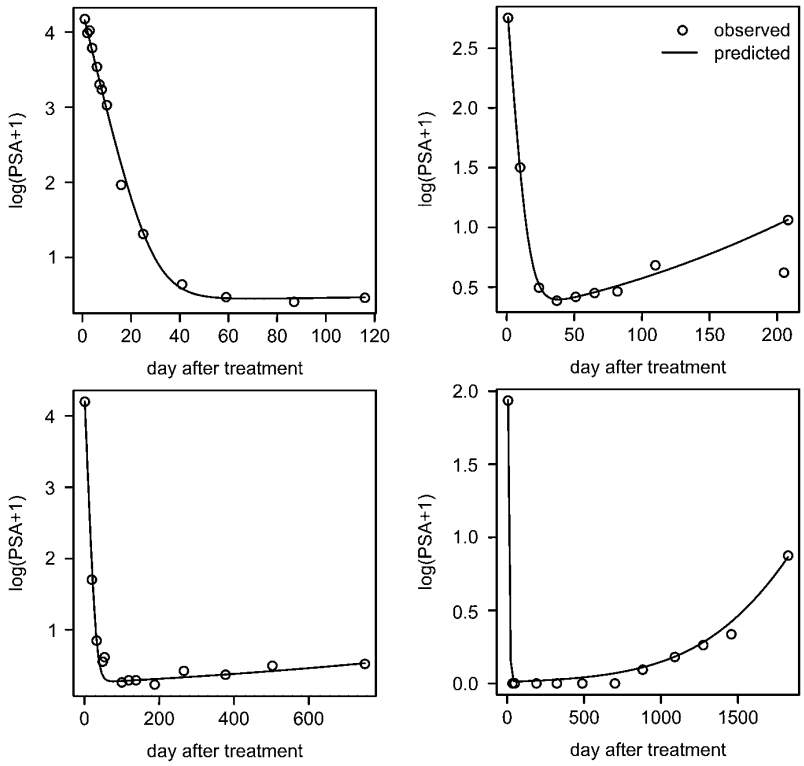

Figure 2. Plot of the logarithm of PSA values according to the day after treatment for four randomly selected patients, and of the trajectory predicted by the Student/Dirichlet model for these patients.

Statistics
in Medicine

Suppose the ith of n patients had m_i longitudinal measurements denoted y_{ij}, $j=1,\ldots,m_i$, taken at time t_{ij} after treatment, j indexing the measurement within that patient data. The longitudinal model is given by:

$$y_{ij} = \exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \tag{2}$$

$$r_{1i} \sim N(\mu_{r_1}, \sigma_{r_1}^2), \quad r_{2i} \sim N(\mu_{r_2}, \sigma_{r_2}^2), \quad r_{3i} \sim N(\mu_{r_3}, \sigma_{r_3}^2), \quad r_{4i} \sim N(\mu_{r_4}, \sigma_{r_4}^2) \tag{3}$$

A normal distribution was assumed for residuals. A comparison of nested models, adding successively each parameter as an additional random effect, indicated that the use of a random effect for each parameter instead of a simple fixed effect increased significantly the likelihood. The covariance between random effects could not be modelled because of few measurements in some patients. More details on the implementation of this model are given in Appendix A.

The predicted values of the random effects were used to calculate the three individual dynamic PSA criteria that may be used for the diagnosis of recurrence:

- the time to the nadir:

$$T_i = \frac{1}{r_{2i}+r_{4i}} \ln\left(\frac{\exp(r_{1i})r_{2i}}{\exp(r_{3i})r_{4i}}\right)$$

- the nadir value: $N_i = \exp(r_{1i})\exp(-r_{2i}T_i) + \exp(r_{3i})\exp(r_{4i}T_i)$;
- the velocity, given directly by r_{4i}.

Unbiased estimates of the random effects are required for a good assessment of the diagnostic accuracy of the dynamic PSA criteria.

2.4. The transform-both-sides methodology

The distribution of all PSA measurements was right-skewed. According to Carroll and Ruppert [17], skewed data may be modelled by transforming both sides of equation (1) with a chosen function that adjusts for the skewness of the distribution. The method of Lipsitz *et al*. [18] was used to choose the most suitable transformation among the Box–Cox family; the estimated value for the parameter of the Box–Cox transformation ranged between -0.5 and 0; it was rounded to 0, which corresponds to the current logarithmic transformation in the PSA studies. A value of one was added to minimize the influence of extremely low PSA values [11]. The final model including the random effect distribution shown in (3) was hence:

$$\ln(y_{ij}+1) = \ln(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \tag{4}$$

This will be thereafter called the Gauss/Gauss model. It reproduced well individual PSA trajectories.

3. Modelling intra- and inter-patient variability

3.1. Intra-patient variability

The plot of the Gauss/Gauss standardized residuals against the predicted values (Figure 3) shows outlying values. These outlying residuals corresponded mainly to the aforementioned occasional PSA fluctuations. Because we were not interested in modelling these fluctuations, a solution would have been to discard these measurements; however, it was sometimes difficult to decide which measurement was due to PSA occasional fluctuations. Because inferences based on normal distribution are known to be vulnerable to outliers, keeping these measurements in the data might have biased the random effect estimates.

Hence, in estimating the parameters, we have decided to give less weight to these measurements by replacing the normal distribution of the residuals by a Student-t [19, 20] distribution, which led to the following model:

$$\ln(y_{ij}+1) = \ln(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim t(\sigma^2, \nu) \tag{5}$$

where t denotes the Student-t scaled distribution, σ the scale parameter, and ν the degrees of freedom. Details on the implementation of the Student-t distribution are given in Appendix A. In this model, the contribution of the measurements to the likelihood is weighted according to the inverse of the square differences between observed and predicted values relative to the residual standard error [19]; this should downweight the influence of the outliers on parameter estimates. The small ν value estimated from the data (1.68 [1.45–1.88]) confirmed the presence of the outlying measurements.

Figure 4 presents the PSA trajectories predicted by the models with Gaussian and Student-t residuals for a patient with an outlying measurement at 236 days. The last model gave less weight to this point; this induced a lower rate of increase, which matched other measurements, and should reduce the bias in the r_4 parameter estimate. A more detailed comparison of parameter estimates will be given later.

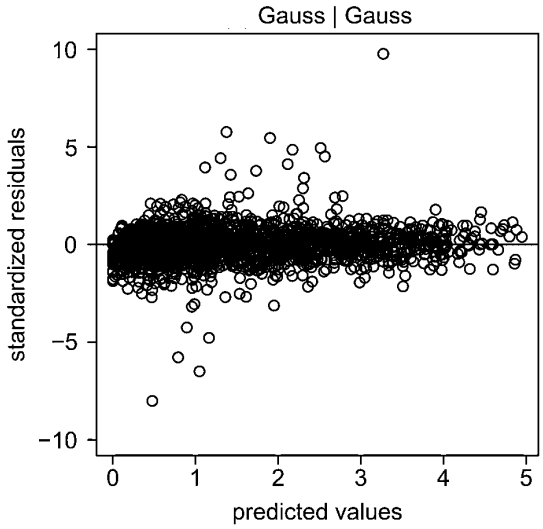

Figure 3. Plot of the standardized residuals against the predicted values for the Gauss/Gauss model.

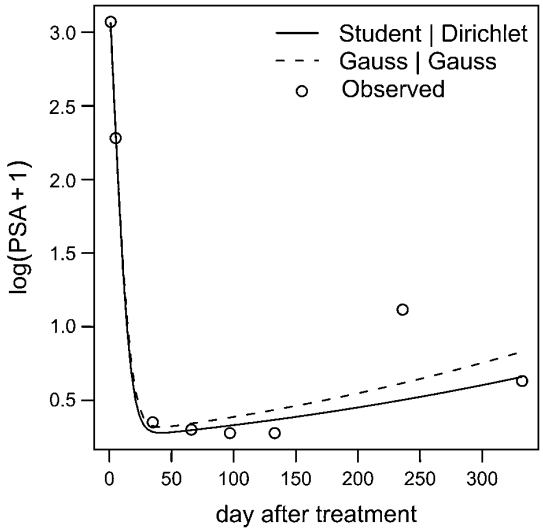

Figure 4. Plot of the logarithm of PSA values for one selected patient: observed, predicted with the Gauss/Gauss model and with the Student/Dirichlet model.

3.2. Inter-patients variability

In the previous model, random effects were supposed to stem from normal distributions. This assumption is quite unrealistic when the population consists of a mixture of sub-populations (at least diseased and not diseased patients). Q-Q plots of the predicted random effects (not shown) rejected the hypothesis of normality for r_1 and r_4, but not for r_2 and r_3; this may bias the random effects estimates [21].

The assumption of normality on r_1 and r_4 distributions was relaxed by considering a Dirichlet process as prior for these distributions. Putting masses on several points belonging to the support of the distribution of the random effects results in estimates stemming from a finite mixture of degenerate distributions. The idea of mixture is appealing in a PSA study because the study population is a mixture of patients at different stages of prostate cancer. More information on Dirichlet processes can be found in two articles by Escobar [22] and Escobar and West [23]. These processes are easily implemented using the Gibbs

Statistics
in Medicine

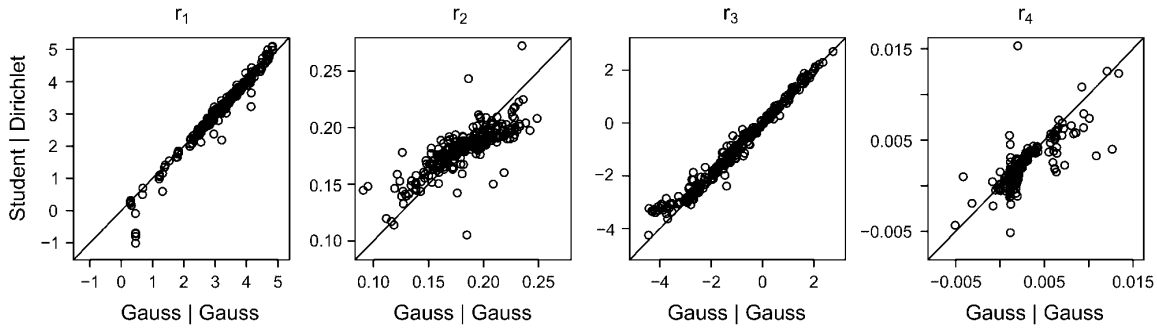

Figure 5. Plot of the Gauss/Gauss random effect estimates against the Student/Dirichlet ones.

sampler (details are given in Appendix A). This led to the Student (residuals)/Dirichlet (random effects) model:

$$\ln(y_{ij}+1) = \ln(\exp(r_{1i})\exp(-r_{2i}t_{ij})+\exp(r_{3i})\exp(r_{4i}t_{ij})+1)+\varepsilon_{ij}, \quad \varepsilon_{ij}\sim t(\sigma^2,\nu)$$

$$r_{1i}\sim \mathrm{DP}(M_1G_{01}), \quad r_{2i}\sim \mathrm{N}(\mu_{r_2},\sigma^2_{r_2}), \quad r_{3i}\sim \mathrm{N}(\mu_{r_3},\sigma^2_{r_3}), \quad r_{4i}\sim \mathrm{DP}(M_4G_{04})$$

$$G_{01}\sim \mathrm{N}(\mu_{r_1},\sigma^2_{r_1}), \quad G_{04}\sim \mathrm{N}(\mu_{r_4},\sigma^2_{r_4})$$

DP denotes a Dirichlet process, G_{01} and G_{04} correspond to the prior expectation on the distributions of r_1 and r_4. The full conditionals of this model are available at: `ftp://pbil.univ-lyon1.fr/pub/datasets/SUBTIL09/index.html`.

3.3. Comparison of random effects estimates

The Gauss/Gauss model described well individual PSA trajectories but some of its assumptions were not fulfilled. The Student/Dirichlet model took into account outlying measurements and relaxed the assumption of normality of random effects. The random effect estimates stemming from this model and from the Gauss/Gauss one differed greatly regarding r_2 and r_4 (Figure 5). As the Student-*t* distribution and Dirichlet process were introduced to take into account the outlying values and the non-normal random effect distributions of the PSA data, the differences in individual parameter estimates might result from bias reduction. This was investigated below in a simulation study.

4. Simulation study

Several patterns of residuals and random effects were simulated to compare the results obtained with the Gauss/Gauss model to those obtained with the Student/Dirichlet one. The Student/Gauss model was not considered as an intermediate model—nor the Gauss/Dirichlet one—because the Student distribution alone would not have corrected the random effect distributions and Dirichlet processes alone would not have downweighted the measurements resulting from the PSA fluctuations.

4.1. Simulation design

Let us consider n patients, of whom $n/2$ are developing a recurrence (cases) and $n/2$ are not (controls). All are going to have a marker measurement just after the treatment, every 5 days during the first 20 days, then every 20 days until the 100th day. As those measurements are often delayed, the actual test times were sampled from a uniform distribution whose left boundary corresponds to the day the measurement should have been done and the right boundary to the day the next measurement should be done. The last day of follow-up was drawn, for each patient, from a Weibull distribution, with shape 9 and scale 100. Biomarker values were generated using a simpler model than the one used to describe PSA trajectories

$$\ln(y_{ij}) = \ln(\exp(r_{1i})+\exp(r_{2i}t_{ij}))+\varepsilon_{ij}$$

$$r_{1i}\sim \mathrm{N}(1.2,0.03^2), \quad r_{2i}\sim \begin{cases} \mathrm{N}(-0.11,0.005^2) & \text{for controls} \\ \mathrm{N}(-0.09,\sigma_c^2) & \text{for cases} \end{cases}$$

$$\varepsilon_{ij}\sim p_0\times \mathrm{N}(0,\sigma^2_{\varepsilon_{ij}})+(1-p_0)\times \mathrm{N}(0,0.02^2), \quad \sigma_{\varepsilon_{ij}}\sim \mathrm{Unif}(0.02,\sigma_u)$$

where p_0 corresponds to the proportion of the outlying measurements. Parameter default values were $\sigma_c=0.01$, $p_0=\frac{1}{7}$, $\sigma_u=0.06$, and $n=100$. A slow decrease in biomarker—i.e. a small value of r_2—is indicative of disease.

Simulations were conducted by varying the form of the r_2 values distribution (σ_c), the proportion of outliers (p_0), the scale of biomarker outliers (σ_u), and the number of patients (n). Each set of simulations involved 100 replicates; the sampling times, the random effect values, and the measurement type—outlier or not—were kept constant over the 100 replicates. At each replicate, the parameters were estimated using the two aforementioned models whose complete specification and priors are described in Appendix B. In each model, the Gibbs sampler algorithm was implemented in R [24] (source codes are available at: `ftp://pbil.univ-lyon1.fr/pub/datasets/SUBTIL09/index.html`).

For r_2, we analysed the relative bias, the width of the 95 per cent confidence interval divided by the true value, and the coverage probability with both models for each patient. The results are summarized using the quartiles of these criteria over all patients and the interquartile range (Tables I, II, and III). The standard deviations of these quartiles were calculated by bootstrap, using 1000 samples from the original 100 replicates. The mean absolute ratio of the bias with the Gauss/Gauss model to the bias with the Student/Dirichlet model and the relative efficiency were also computed for each patient. When θ is the true r_2 value for a patient and $\hat{\theta}_1$ and $\hat{\theta}_2$ are, respectively, the Gauss/Gauss and the Student/Dirichlet estimates, the relative efficiency is defined as $E[(\hat{\theta}_1-\theta)^2]/E[(\hat{\theta}_2-\theta)^2]$.

4.2. Results

For each simulation, the relative bias in r_2 over all patients was approximately centred to zero, with both models, with a little shift in general with the Gauss/Gauss model. The interquartile range of relative bias was generally higher with the Gauss/Gauss model than with the Student/Dirichlet model, with differences from 0.02 to 0.08; standard deviations were small, indicating that these differences were not due to simulation-based variability. Moreover, the quartiles of the relative bias were in most cases higher with the Gauss/Gauss model than with the Student/Dirichlet one. The median over all patients of the mean of the ranges of the 95 per cent confidence intervals of the estimated r_2 values divided each by the corresponding true value was generally slightly higher with the Student/Dirichlet model than with the Gauss/Gauss model, but the interquartile range was smaller. The coverage probability was generally around 0.96 with the Gauss/Gauss model and around 0.92 with the Student/Dirichlet model (Tables I, II, and III).

The first quartile of the means of the absolute ratios of the bias with the Gauss/Gauss model to the bias with the Student/Dirichlet model varied from 2.03 to 2.99, the median varied from 3.05 to 4.26, and the third quartile from 4.79 to 6.59. The first quartile of the relative efficiencies was almost always above 1 (data not shown), which was in favour of the Student/Dirichlet model.

The distribution of r_2 values for $\sigma_c=0.005$ was a mixture of two well-separated normal distributions that got closer as σ_c increased. The difference between the interquartile ranges obtained with the Gauss/Gauss model and the Student/Dirichlet model increased, whereas σ_c decreased (Table I); that is, the r_2 distribution was less easily reproduced by a normal distribution. This reflects the ability of the Dirichlet process prior to limit the bias in random effect estimates when they are not issued from a normal distribution. The aforementioned difference increased with the increasing proportion or the scale of outlying measurements (Tables II and III); this reflects the ability of the Student-t distribution to limit the bias in random effect estimates when there are outlying values in biomarker measurements. Increasing the number of patients from $n=50$ to $n=200$ did not change the difference between the interquartile ranges calculated according to the Gauss/Gauss model and to the Student/Dirichlet model (data not shown).

5. The diagnostic accuracy of dynamic PSA criteria

The random effect estimates stemming from the models (Gauss/Gauss or Student/Dirichlet) were used to calculate three dynamic PSA criteria for each patient: the nadir, the time to this nadir, and the velocity. The ability of these criteria to make the diagnosis of recurrence of prostate cancer was assessed in terms of sensitivity (true-positive rate) and specificity (one minus the false-positive rate). As these criteria were quantitative, their overall diagnostic accuracy was summarized by the area under the receiver operating characteristic (ROC) curve, the plot of sensitivity against one minus the specificity across all possible thresholds values set for the test; higher values indicate better diagnostic test [25].

This section investigates the impact of the choice of the model on the estimation of the diagnostic test accuracy.

5.1. Results of the simulation study

In the previous simulation study, r_2 values may be used as diagnostic test. As their theoretical distribution was known in diseased and non-diseased patients, the true AUC could be assessed and compared with the estimated ones stemming from the results of the Gauss/Gauss and the Student/Dirichlet models. The relative bias of the AUC estimates and their 95 per cent confidence intervals divided by the true value over the 100 replicates are reported (Table IV), as well as the standard deviations of these criteria, calculated using bootstrap (1000 samples from the 100 replicates). The mean absolute ratio of the bias with the Gauss/Gauss model to the Student/Dirichlet model was also computed.

For each simulation, the relative bias for AUC estimates was almost always smaller with the Student/Dirichlet model than with the Gauss/Gauss model with differences reaching 0.06. The mean absolute ratio of the bias with the Gauss/Gauss model to the bias with the Student/Dirichlet model varied from 1.13 to 1.47, the relative efficiencies were always above 1 (data not shown).

Statistics
in Medicine

Table I. Simulation study—comparison between the Gauss/Gauss and the Student/Dirichlet results varying the form of the r_2 values distribution (σ_c), with $p_0 = \frac{1}{7}$, $\sigma_u = 0.06$, and $n = 100$. Values in brackets denote standard deviations.

	r_2 relative bias*				r_2 coefficient of variation*†				r_2 coverage probability*			
	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡
$\sigma_c = 0.005$												
Gauss/Gauss	−0.0774	0.0012	0.0616	0.1390	0.40	0.44	0.48	0.08	0.94	0.97	0.99	0.05
	(0.0042)	(0.0046)	(0.0033)	(0.0045)	(0.004)	(0.004)	(0.004)	(0.003)	(0.006)	(0.005)	(0.004)	(0.007)
Student/Dirichlet	−0.0276	0.0101	0.0437	0.0713	0.48	0.50	0.52	0.04	0.91	0.94	0.95	0.04
	(0.0036)	(0.0039)	(0.0039)	(0.0048)	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.005)	(0.004)	(0.007)
$\sigma_c = 0.01$												
Gauss/Gauss	−0.0451	0.0229	0.0584	0.1035	0.47	0.50	0.59	0.12	0.92	0.96	0.98	0.06
	(0.0048)	(0.0040)	(0.0037)	(0.0055)	(0.003)	(0.003)	(0.006)	(0.005)	(0.007)	(0.005)	(0.004)	(0.007)
Student/Dirichlet	−0.0326	0.0061	0.0375	0.0701	0.49	0.51	0.55	0.06	0.89	0.92	0.94	0.05
	(0.0042)	(0.0036)	(0.0038)	(0.0054)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.007)
$\sigma_c = 0.02$												
Gauss/Gauss	−0.0512	0.0174	0.0591	0.1103	0.47	0.49	0.56	0.09	0.91	0.96	0.98	0.07
	(0.0061)	(0.0039)	(0.0034)	(0.0067)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)	(0.007)
Student/Dirichlet	−0.0380	0.0049	0.0402	0.0782	0.47	0.50	0.55	0.08	0.89	0.92	0.95	0.06
	(0.0038)	(0.0035)	(0.0037)	(0.0048)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)	(0.007)

*Mean value for each patient over the 100 replicates described by the quartiles and interquartile range over all patients.
†Coefficient of variation: length of the 95 per cent confidence interval divided by the true value.
‡Interquartile range.

Statistics in Medicine

Table II. Simulation study—comparison between the Gauss/Gauss and the Student/Dirichlet results varying the proportion of outliers (p_0), with $\sigma_c = 0.01$, $\sigma_u = 0.06$, and $n = 100$. Values in brackets denote standard deviations.

	r_2 relative bias*				r_2 coefficient of variation*†				r_2 coverage probability*			
	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡
$p_0 = \frac{1}{14}$												
Gauss/Gauss	−0.0369	0.0196	0.0525	0.0894	0.46	0.48	0.57	0.11	0.92	0.96	0.98	0.06
	(0.0051)	(0.0031)	(0.0036)	(0.0056)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.008)
Student/Dirichlet	−0.0229	0.0016	0.0421	0.0650	0.46	0.49	0.53	0.07	0.89	0.91	0.94	0.05
	(0.0043)	(0.0031)	(0.0037)	(0.0049)	(0.003)	(0.003)	(0.004)	(0.004)	(0.007)	(0.005)	(0.005)	(0.008)
$p_0 = \frac{1}{7}$												
Gauss/Gauss	−0.0451	0.0229	0.0584	0.1035	0.47	0.50	0.59	0.12	0.92	0.96	0.98	0.06
	(0.0048)	(0.0040)	(0.0037)	(0.0055)	(0.003)	(0.003)	(0.006)	(0.005)	(0.007)	(0.005)	(0.004)	(0.007)
Student/Dirichlet	−0.0326	0.0061	0.0375	0.0701	0.49	0.51	0.55	0.06	0.89	0.92	0.94	0.05
	(0.0042)	(0.0036)	(0.0038)	(0.0054)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.007)
$p_0 = \frac{2}{7}$												
Gauss/Gauss	−0.0548	0.0252	0.0596	0.1144	0.51	0.54	0.64	0.13	0.92	0.95	0.98	0.06
	(0.0065)	(0.0039)	(0.0040)	(0.0069)	(0.004)	(0.004)	(0.008)	(0.007)	(0.007)	(0.005)	(0.005)	(0.008)
Student/Dirichlet	−0.0379	−0.0011	0.0415	0.0794	0.54	0.58	0.63	0.09	0.89	0.93	0.95	0.06
	(0.0048)	(0.0048)	(0.0042)	(0.0059)	(0.004)	(0.004)	(0.005)	(0.005)	(0.007)	(0.006)	(0.005)	(0.008)

*Mean value for each patient over the 100 replicates described by the quartiles and interquartile range over all patients.
†Coefficient of variation: length of the 95 per cent confidence interval divided by the true value.
‡Interquartile range.

Statistics
in Medicine

Table III. Simulation study—comparison between the Gauss/Gauss and Student/Dirichlet results varying the scale of biomarker outliers (σ_U), with $\sigma_C = 0.01$, $p_0 = \frac{1}{7}$, and $n = 100$. Values in brackets denote standard deviations.

	r_2 relative bias*				r_2 coefficient of variation*†				r_2 coverage probability*			
	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡	1st Q	2nd Q	3rd Q	IQR‡
$\sigma_U = 0.03$												
Gauss/Gauss	−0.0434	0.0134	0.0497	0.0931	0.44	0.47	0.54	0.10	0.92	0.95	0.97	0.05
	(0.0056)	(0.0033)	(0.0032)	(0.0057)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.007)
Student/Dirichlet	−0.0335	0.0047	0.0400	0.0735	0.44	0.47	0.52	0.08	0.89	0.91	0.93	0.04
	(0.0040)	(0.0036)	(0.0037)	(0.0049)	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.005)	(0.005)	(0.007)
$\sigma_U = 0.06$												
Gauss/Gauss	−0.0451	0.0229	0.0584	0.1035	0.47	0.50	0.59	0.12	0.92	0.96	0.98	0.06
	(0.0048)	(0.0040)	(0.0037)	(0.0055)	(0.003)	(0.003)	(0.006)	(0.005)	(0.007)	(0.005)	(0.004)	(0.007)
Student/Dirichlet	−0.0326	0.0061	0.0375	0.0701	0.49	0.51	0.55	0.06	0.89	0.92	0.94	0.05
	(0.0042)	(0.0036)	(0.0038)	(0.0054)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.007)
$\sigma_U = 0.12$												
Gauss/Gauss	−0.0598	0.0348	0.0809	0.1407	0.52	0.54	0.65	0.13	0.91	0.96	1.00	0.09
	(0.0093)	(0.0041)	(0.0043)	(0.0098)	(0.004)	(0.004)	(0.008)	(0.007)	(0.006)	(0.006)	(0.005)	(0.008)
Student/Dirichlet	−0.0332	0.0061	0.0456	0.0788	0.51	0.53	0.59	0.08	0.88	0.92	0.94	0.06
	(0.0045)	(0.0037)	(0.0042)	(0.0060)	(0.003)	(0.004)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.008)

*Mean value for each patient over the 100 replicates described by the quartiles and interquartile range over all patients.
†Coefficient of variation: length of the 95 per cent confidence interval divided by the true value.
‡Interquartile range.

Table IV. Simulation study—comparison of the Gauss/Gauss and Student/Dirichlet results according to AUC values. Values in brackets denote standard deviations.

	Gauss/Gauss		Student/Dirichlet	
	Relative bias	CV*	Relative bias	CV*
$\sigma_c = 0.005^{\dagger}$	−0.180 (0.005)	0.058 (0.009)	−0.140 (0.005)	0.066 (0.009)
$\sigma_c = 0.01^{\dagger}$	−0.185 (0.004)	0.054 (0.008)	−0.161 (0.004)	0.052 (0.008)
$\sigma_c = 0.02^{\dagger}$	−0.231 (0.005)	0.069 (0.009)	−0.201 (0.005)	0.056 (0.009)
$p_0 = \frac{1}{14}^{\ddagger}$	−0.171 (0.004)	0.058 (0.009)	−0.155 (0.004)	0.053 (0.009)
$p_0 = \frac{1}{7}^{\ddagger}$	−0.185 (0.004)	0.054 (0.008)	−0.161 (0.004)	0.052 (0.008)
$p_0 = \frac{2}{7}^{\ddagger}$	−0.210 (0.005)	0.055 (0.008)	−0.189 (0.005)	0.062 (0.008)
$\sigma_u = 0.03^{\S}$	−0.162 (0.004)	0.049 (0.007)	−0.161 (0.005)	0.065 (0.007)
$\sigma_u = 0.06^{\S}$	−0.185 (0.004)	0.054 (0.008)	−0.161 (0.004)	0.052 (0.008)
$\sigma_u = 0.12^{\S}$	−0.246 (0.005)	0.065 (0.007)	−0.186 (0.005)	0.064 (0.007)

*Coefficient of variation: length of the 95 per cent confidence interval divided by the true value.
†The other parameters were $p_0 = \frac{1}{7}$, $\sigma_u = 0.06$, and $n = 100$.
‡The other parameters were $\sigma_c = 0.01$, $\sigma_u = 0.06$, and $n = 100$.
§The other parameters were $\sigma_c = 0.01$, $p_0 = \frac{1}{7}$, and $n = 100$.

Table V. Clinical example—AUC estimates [95 per cent confidence interval] stemming from the Gauss/Gauss model, the Student/Gauss model, and the Student/Dirichlet model in four dynamic PSA criteria.

	Gauss/Gauss	Student/Dirichlet
Nadir	0.685 [0.672, 0.695]	0.691 [0.684, 0.698]
Time to the Nadir	0.630 [0.583, 0.676]	0.645 [0.608, 0.678]
Velocity	0.561 [0.517, 0.633]	0.538 [0.487, 0.588]
Nadir+Time to the Nadir	0.687 [0.672, 0.713]	0.705 [0.683, 0.717]

5.2. Results of the PSA study

Each iteration of the Gibbs sampler gave an estimate of r_1, r_2, r_3, and r_4 for each patient; thus, an estimate of the dynamic PSA criteria. At each iteration, the diagnostic accuracy of these dynamic PSA criteria was assessed by building the corresponding ROC curve and estimating the AUC. The posterior distribution of the AUCs for each diagnostic test was derived from the results of all retained iterations (Table V). This process took into account the variability due to the uncertainty around the estimates of the PSA dynamic criteria.

Overall, the results given by the two models were comparable, with slightly higher AUC estimates and tighter confidence intervals with the Student/Dirichlet model versus the Gauss/Gauss model, except for velocity.

Whatever the model, velocity seemed to be a poor diagnostic test (AUC values around 0.5), whereas the nadir value was the best one. Other dynamic PSA criteria, such as the rate of decrease, were analysed, but were not found linked to recurrence. The combination of the nadir and the time to the nadir using a logistic regression led to a more accurate diagnostic test than the nadir alone, but the difference was not statistically significant.

6. Discussion

This article presented a longitudinal model for PSA measurements after an HIFU treatment. We believe this model can be also used after other treatments, such as radiotherapy, because PSA curves are similar. It was emphasized that residuals and random effects may not have normal distributions in PSA measurements. Simulation results showed that this can bias random effects estimates if not taken into account, and thus, leads to a poor modelling of the measurements.

The Student/Dirichlet model, proposed to relax the normal distribution hypotheses, was shown to give in general less biased random effect estimates than the Gauss/Gauss one. Whatever the set of simulations, the first quartile of the mean absolute ratio of the biases of the two models over all patients was always higher than 2, which is greatly in favour of the Student/Dirichlet model. This model seems overall more efficient than the Gauss/Gauss one. The coverage probability was somewhat low with the Student/Dirichlet method and somewhat high with the Gauss/Gauss method, but close to 0.95 with both. Those results support the use of the Student/Dirichlet model in this specific PSA study.

Statistics
in Medicine

We do not claim that the proposed exponential decay–exponential growth model is always the best one to describe the PSA measurements after a treatment; however, it seemed to be well suited to our data and led to simple calculations of the nadir, the time to nadir, and the velocity. The Student-*t* distribution and Dirichlet process have been used to relax hypotheses that were too restrictive for PSA data though other solutions may be considered. To our knowledge, except Proust-Lima and Taylor [26] who used a mixture of normals for random effects, little attention has been paid to these hypotheses. In our application, Dirichlet processes have been preferred to the Student-*t* distribution or to skewed extensions of the normal and Student-*t* distributions because these processes were thought more flexible and less restrictive when there is no prior knowledge about the distributions of the random effects. Their major drawback is to generate discrete distributions with probability one [27]. The Polya trees [28] would be a good alternative to generate continuous distributions with probability one under some conditions, but with computational complications. The Student-*t* distribution has been already used for residuals in linear mixed models for robust inferences [29], but, to our knowledge, not in non-linear ones. Some other distributions could have been used instead, such as the contaminated normal and the slash distributions [20]; the choice depends partly on the proportion of outliers and on how far these are from the 'center' of the distribution. Nonetheless, the Student-*t* distribution is always a good alternative to the Gaussian distribution and is, besides, easy to implement; it gave satisfactory results in our study.

In the exponential decay–exponential growth model, all individual parameters $\{r_1, r_2, r_3, r_4\}$ were assumed as random effects because, in a nested model strategy, considering each of these parameters as a random effect instead of a simple fixed effect increased significantly the model likelihood.

Simulation results also emphasized the impact of a bias in the estimates of a dynamic criterion on the estimated diagnostic accuracy. Random errors in marker measurements are known to bias AUC estimates towards 0.5, meaning poor diagnostic test [30]. With less biased estimates of r_2 in the simulation part with the Student/Dirichlet model, AUC values tended to be higher and closer to the true ones. In the application part, however, AUC values were similar whatever the model (Student/Dirichlet or Gauss/Gauss). Because the random effect estimates in the two models were quite different, we expected more differences in AUC, but a change in the random effect estimates might not change the relative order of the dynamic PSA criteria between diseased and non-diseased patients, and consequently, AUC values. Anyway, we are more confident in AUC estimates stemming from the Student/Dirichlet model than from the Gauss/Gauss one because the former rely on less biased random effect estimates.

Another area of concern is the method used to sample from posterior distributions of random effects that have no explicit forms. Adaptative rejection Metropolis sampling was used by Yu *et al.* [11] but, in our case, this method prevented the random parameter estimates from converging. A Metropolis algorithm was used instead, with a jumping distribution scale chosen to obtain an acceptance rate close to 0.23, as recommended by Gelman *et al.* [9]; however, we did not analyse the impact of the sampling method on the random effect estimates, and, consequently, on the AUC estimates.

In the clinical study, the nadir, or a combination of the nadir and the time to the nadir, were found to be the best dynamic criteria stemming from PSA trajectories for the diagnosis of recurrence of prostate cancer, with AUC values around 0.7. Uchida *et al.* [4] have also underlined the association between the nadir and the recurrence of prostate cancer but they have calculated the nadir directly from PSA measurements. Blana *et al.* [5] have proposed a diagnostic test based on PSA crossing a threshold defined by the nadir plus 1.2 ng/mL, whatever the nadir value. The use of the nadir value itself leads however to earlier positive diagnostic tests. Using this dynamic PSA criterion as a diagnostic test requires the determination of a nadir threshold. In our study, as in the one by Uchida *et al.* [4], the gold standard was defined on biopsy results; however, we have considered the results of all biopsies taken after 3 months and not only the result at 6 months. Biospsy-based diagnostic tests are known to lack sensitivity; the use of several biopsies might partly compensate that lack. One limit of the study is that we have considered only patients with at least one biopsy after 3 months, thus already suspected of developing recurrence. Hence, our results might not be applicable to the whole population of patients who underwent an HIFU treatment.

In the field of prostate cancer, we believe that the use of the Student-*t* distribution for residuals and Dirichlet process as prior for random effects—or other alternatives to the classical Gauss/Gauss model—should be generalized to achieve a better modelling of individual PSA trajectories. Here, we focused only on the assessment of PSA diagnostic accuracy, but the proposed methods would also be of interest in the analysis of the PSA prognostic value [11] or the comparison of treatment effects. In some trials, comparing several doses of external beam radiation for the treatment of localized prostate cancer, less success was considered linked to an increase in raw PSA measurements [31, 32]. A comparison robust to measurements errors and occasional fluctuations would be obtained using modelled PSA trajectories. Several recommendations have been made by the PCWG2 group [33] for the evaluation of systemic treatments in prostate cancer, especially on eligibility criteria and outcome measures. By carefully modelling PSA trajectories, the present work is an important step to improve the quality of that evaluation.

Generally, in longitudinal models, when one is interested not only in the fixed effect estimates but also in the random ones, more attention should be paid to the distribution of random effects and, depending on the data, to the distribution of residuals. The Student-*t* distribution is always a good alternative to the Gaussian one, and is, besides, easy to implement. For random effects, several alternatives are possible; Dirichlet processes seem to offer a very flexible one.

Appendix A: Models implementation

A.1. The Gauss/Gauss model

A.1.1. Implementation. The model was fitted using the Gibbs sampler algorithm [9], a Markov Chain Monte Carlo Method. Explicit forms of posterior distributions existed for all parameters, except for random effects (because of the non-linearity of the model).

They were consequently sampled using the Metropolis algorithm with normal proposal. The algorithm was implemented in R [24], but could also be easily implemented in WINBUGS [34]. The results were based on 1 in 100 thinning on the total 300 000 iterations, after a burn-in period of 50 000 iterations.

A.1.2. Priors. Uninformative normal hyperpriors were used for the means and Gamma hyperpriors for the inverse of variance parameters.

A.2. The Student/Dirichlet model

A.2.1. The Student-t distribution for residuals. A simple way to use the Student-*t* distribution in a Bayesian approach is to make use of its equivalence with an appropriate mixture of normal distributions [35]:

$$y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim t(\sigma^2, \nu) \Leftrightarrow y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, V_i), \quad V_i \sim \mathrm{Inv}-\chi^2(\nu, \sigma^2)$$

Inv-χ^2 denotes the scaled inverse χ^2 distribution and V_i parameters may be seen as auxiliary parameters introduced only to simplify the model computation. Conditional on data ($\boldsymbol{y}$) and on other parameters of the model (*rest*), the V_i are independent and have an inverse χ^2 posterior distribution [9]. If a uniform prior is assumed for $\ln(\sigma)$, then σ^2 has a Gamma posterior distribution [9]. For the inverse of the degrees of freedom ν, it is common to assume a uniform prior within the range [0, 1]. There is no simple way to sample directly from the posterior induced by this prior. A Metropolis algorithm with normal proposal was consequently used, as proposed by Gelman *et al.* [9].

A.2.2. Dirichlet process prior for random effects. When a Dirichlet process is used as prior for the distribution of a random effect on patients, the value for each patient is sampled sequentially. Let β_{-i} denote the set of random effect estimates for all patients except the ith one (β could correspond to r_1 or r_4 in the PSA study). As derived in Escobar [22], conditional on β_{-i} and on other parameters, the full conditional posterior distribution of β_i is given by

$$p(\beta_i | \beta_{-i}, \boldsymbol{y}, \text{rest}) \propto \sum_{\substack{k=1 \\ k \neq i}}^{n} q_k \delta(\beta_k) + q_0 G_i(\beta_i) \tag{A1}$$

where $G_i(\beta_i) = p(\beta_i | \boldsymbol{y}_i) \propto G_0(\beta_i) \times p(\boldsymbol{y}_i | \beta_i)$ and $\delta(\beta_k)$ represents an indicator function at β_k. The weights q_k and q_0 are defined by:

$$q_k \propto p(\boldsymbol{y}_i | \beta_k), \quad q_0 \propto M \int \cdots \int p(\boldsymbol{y}_i | \beta_i) \, \mathrm{d}G_0(\beta_i)$$

$G_i(\beta_i)$ is the posterior distribution of $\beta_i | \boldsymbol{y}_i$ under a prior G_0; the weight q_0 is proportional to M times the marginal density of $\boldsymbol{y}_i$ using G_0 as a prior for β_i; the weight q_k is proportional to the likelihood of data $\boldsymbol{y}_i$ evaluated with the trajectory parameter vector of subject k. Hence, the conditional distribution $p(\beta_i | \beta_{-i}, \boldsymbol{y}, \text{rest})$ is a weighted mixture of the best guess of the prior G_0 and of mass point distributions associated with β_{-i}. M is a parameter that estimates the amount of mixing in the model and measures the variability of G around G_0; G tends to G_0 as $M \to \infty$; as $M \to 0$, all β_i vectors tend to the same vector; the model may then be seen as a fixed effect model. A Gamma prior was used for M, as suggested by Escobar and West [36].

It is straightforward to sample vectors β_i from equation (A1) when q_0 has a closed-form solution. However, in models associated with a longitudinal marker, it is often impossible to find G_0 conjugated to the likelihood. In those situations, Neel [37] proposed several algorithms to sample from the conditional posterior of vectors β_i. The present article used Neal algorithm 8 that relies on the use of auxiliary variables.

A.2.3. Implementation. The model was implemented in R. The results were based on 1 in 25 thinning on the total 75 000 iterations, after a burn-in period of 150 000 iterations.

A.2.4. Priors. Uninformative normal hyperpriors were used for means and Gamma hyperpriors for the inverse of variance parameters; uniform priors were assumed for $1/\nu$ and $\ln(\sigma)$. Gamma(1, 1) priors were used for M_1 and M_4; we also performed the analysis with a Gamma(0.1, 0.1) and a Gamma(10, 10) prior, but they led to the same results in terms of random effect estimates.

A.3. Convergence checking

Two parallel chains of equal lengths with different initial values were run for each model. The estimate of each parameter was consistent across the chains, as judged by the Gelman–Rubin diagnostic. The thinning parameter was chosen to limit autocorrelation.

Appendix B: Specification of the models and priors used in the simulation study

The general model was:

$$\ln(y_{ij}) = \ln(\exp(r_{1i}) + \exp(r_{2i} t_{ij})) + \varepsilon_{ij}$$

Statistics
in Medicine

B.1. The Gauss/Gauss model

$$r_{1i} \sim \mathrm{N}(\mu_{r_1}, \sigma^2_{r_1}), \quad r_{2i} \sim \mathrm{N}(\mu_{r_2}, \sigma^2_{r_2}), \quad \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2_{\varepsilon})$$

Uninformative normal hyperpriors were used for μ_{r_1} and μ_{r_2}, and Gamma hyperpriors for $1/\sigma^2_{r_1}$, $1/\sigma^2_{r_2}$, and $1/\sigma^2_{\varepsilon}$.

B.2. The Student/Dirichlet model

$$r_{1i} \sim \mathrm{N}(\mu_{r_1}, \sigma^2_{r_1}), \quad r_{2i} \sim \mathrm{DP}(MG_0), \quad G_0 \sim \mathrm{N}(\mu_{r_2}, \sigma^2_{r_2}), \quad \varepsilon_{ij} \sim \mathrm{N}(0, V_{ij}), \quad V_{ij} \sim \mathrm{Inv}{-}\chi^2(\nu, \sigma^2)$$

A Gamma(1, 1) was used as prior for M, uninformative normal hyperpriors were used for means and Gamma hyperpriors for the inverse of variance parameters, $1/\nu$ and $\ln(\sigma)$ had the same priors as in the Student/Gauss model.

B.3. Parameter estimation

The results were based on 1 in 4 thinning on the total 12 000 iterations, after a burn-in period of 10 000 iterations for each model. Parameter estimates and confidence intervals were obtained using the half-range mode [38] and the highest posterior density interval [9].

Acknowledgements

This study was partially funded by the French 'Ligue contre le Cancer'. The authors are grateful to Dr Albert Gelet and the members of the 'Service d'Urologie et Chirurgie de la Transplantation' of the 'Hospices Civils de Lyon' (France) for providing the data on PSA after HIFU and introducing them to this clinical problematic. We also thank Pr René Echochard and Dr Jean Iwaz whose suggestions significantly improved the manuscript.

References

1. Scher HI, Warren M, Heller G. The association between measures of progression and survival in castrate-metastatic prostate cancer. *Clinical Cancer Research* 2007; **13**(5):1488–1492. DOI: 10.1158/1078-0432.CCR-06-1885.
2. D'Amico AV, Moul J, Carroll PR, Sun L, Lubeck D, Chen M-H. Prostate specific antigen doubling time as a surrogate end point for prostate cancer specific mortality following radical prostatectomy or radiation therapy. *The Journal of Urology* 2004; **172**(5):S42–S47. DOI: 10.1097/01.ju.0000141845.99899.12.
3. Roach M, Hanks G, Thames H, Schellhammer P, Shipley WU, Sokol GH, Sandler H. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO phoenix consensus conference. *International Journal of Radiation Oncology Biology Physics* 2006; **65**(4):965–974. DOI: 10.1016/j.ijrobp.2006.04.029.
4. Uchida T, Illing RO, Cathcart PJ, Emberton M. To what extent does the prostate-specific antigen nadir predict subsequent treatment failure after transrectal high-intensity focused ultrasound therapy for presumed localized adenocarcinoma of the prostate? *BJU National* 2006; **98**(3):537–539.
5. Blana A, Brown SC, Chaussy C, Conti GN, Eastham JA, Ganzer R, Murat FJ, Pasticier G, Rebillard X, Rewcastle JC, Robertson CN, Thuroff S, Ward JF. High-intensity focused ultrasound for prostate cancer: comparative definitions of biochemical failure. *BJU International* 2009. (Advance Access April 17, 2009.); DOI: 10.1111/j.1464-410X.2009.08518.x.
6. Poissonnier L, Chapelon JY, Rouviere O, Curiel L, Bouvier R, Martin X, Dubernard JM, Gelet A. Control of prostate cancer by transrectal HIFU in 227 patients. *European Urology* 2007; **51**(2):381–387. DOI: 10.1016/j.eururo.2006.04.012.
7. Eastham JA, Riedel E, Scardino PT, Shike M, Fleisher M, Schatzkin A, Lanza E, Latkany L, Begg CB. Variation of serum prostate-specific antigen levels: an evaluation of year-to-year fluctuations. *The Journal of the American Medical Association* 2003; **289**(20):2695–2700.
8. Soletormos G, Semjonow A, Sibley PEC, Lamerz R, Petersen PH, Albrecht W, Bialk P, Gion M, Junker F, Schmid H-P, Van Poppel H, On behalf of the European group on Tumor M. Biological variation of total prostate-specific antigen: a survey of published estimates and consequences for clinical practice. *Clinical Chemistry* 2005; **51**(8):1342–1351. DOI: 10.1373/clinchem.2004.046086.
9. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (2nd edn). Chapman & Hall, CRC: London, 2004.
10. Zagars GK, Pollack A. The fall and rise of prostate-specific antigen. Kinetics of serum prostate-specific antigen levels after radiation therapy for prostate cancer. *Cancer* 1993; **72**(3):832–842. DOI: 10.1002/1097-0142(19930801)72:3<832::AID-CNCR2820720332>3.0.CO;2-6.
11. Yu M, Law NJ, Taylor JMG, Sandler HM. Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* 2004; **14**:835–862.
12. Vollmer RT, Montana GS. The dynamics of prostate-specific antigen after definitive radiation therapy for prostate cancer. *Clinical Cancer Research* 1999; **5**(12):4119–4125.
13. Taylor JMG, Yu M, Sandler HM. Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* 2005; **23**(4):816–825.
14. You B, Perrin P, Freyer G, Ruffion A, Tranchand B, Hénin E, Paparel P, Ribba B, Devonec M, Falandry C, Fournel C, Tod M, Girard P. Advantages of prostate-specific antigen (PSA) clearance model over simple PSA half-life computation to describe PSA decrease after prostate adenomectomy. *Clinical Biochemistry* 2008; **41**:785–795. DOI: 10.1016/j.clinbiochem.2008.04.001.
15. Bellera CA, Hanley JA, Joseph L, Albertsen PC. Hierarchical changepoint models for biochemical markers illustrated by tracking postradiotherapy prostate-specific antigen series in men with prostate cancer. *Annals of Epidemiology* 2008; **18**(4):270–282.
16. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**(4):963–974.
17. Carroll RJ, Ruppert D. *Transformation and Weighting in Regression*. Chapman & Hall: London, 1988.

18. Lipsitz SR, Ibrahim J, Molenberghs G. Using a Box–Cox transformation in the analysis of longitudinal data with incomplete responses. *Journal of the Royal Statistical Society. Series C, Applied Statistics* 2000; **49**(3):287–296. DOI: 10.1111/1467-9876.00192.
19. Lange KL, Little RJA, Taylor JMG. Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association* 1989; **84**:881–896.
20. Rosa GJM, Padovani CR, Gianola D. Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* 2003; **45**(5):573–590. DOI: 10.1002/bimj.200390034.
21. Kleinman KP, Ibrahim JG. A semiparametric Bayesian approach to the random effects model. *Biometrics* 1998; **54**(3):921–938.
22. Escobar MD. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* 1994; **89**(425):268–277.
23. Escobar MD, West M. Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Dey D, Müller P, Sinha D (eds). Springer: New York, 1998; 1–16.
24. R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2008.
25. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.
26. Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* 2009; **10**(3):535–549. DOI: 10.1093/biostatistics/kxp009.
27. Müller P, Quintana FA. Nonparametric Bayesian data analysis. *Statistical Science* 2004; **19**(1):95–110. DOI: 10.1214/088342304000000017.
28. Lavine M. Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics* 1992; **20**:1222–1235. DOI: 10.1214/aos/1176348767.
29. Pinheiro JC, Liu C, Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *Journal of Computational and Graphical Statistics* 2001; **10**(2):249–276.
30. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistics in Medicine* 2000; **19**(16):2115–2129. DOI: 10.1002/1097-0258(20000830)19:16<2115::AID-SIM529>3.0.CO;2-M.
31. Zietman AL, DeSilvio ML, Slater JD, Rossi Jr CJ, Miller DW, Adams JA, Shipley WU. Comparison of conventional-dose vs high-dose conformal radiation therapy in clinically localized adenocarcinoma of the prostate: a randomized controlled trial. *Journal of the American Medical Association* 2005; **294**(10):1233–1239. DOI: 10.1001/jama.294.10.1233.
32. Peeters ST, Heemsbergen WD, Koper PC, van Putten WL, Slot A, Dielwart MF, Bonfrer JM, Incrocci L, Lebesque JV. Dose–response in radiotherapy for localized prostate cancer: results of the Dutch multicenter randomized phase III trial comparing 68 Gy of radiotherapy with 78 Gy. *Journal of Clinical Oncology* 2006; **24**(13):1990–1996. DOI: 10.1200/JCO.2005.05.2530.
33. Scher HI, Halabi S, Tannock I, Morris M, Sternberg CN, Carducci MA, Eisenberger MA, Higano C, Bubley GJ, Dreicer R, Petrylak D, Kantoff P, Basch E, Kelly WK, Figg WD, Small EJ, Beer TM, Wilding G, Martin A, Hussain M, PCCTW Group. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the prostate cancer clinical trials working group. *Journal of Clinical Oncology* 2008; **26**(7):1148–1159. DOI: 10.1200/JCO.2007.12.4487.
34. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337. DOI: 10.1023/A:1008929526011.
35. Geweke J. Bayesian treatment of the independent student-*t* linear model. *Journal of Applied Econometrics* 1993; **8**(S1):19–40.
36. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**(430):577–588.
37. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2000; **9**:249–265.
38. Bickel DR. Robust estimators of the mode and skewness of continuous data. *Computational Statistics and Data Analysis* 2002; **39**(2):153–163. DOI: 10.1016/S0167-9473(01)00057-3.

4.2.2 Principaux résultats de l'article

4.2.2.1 Estimation des effets aléatoires

Les résultats des simulations ont montré qu'en présence de valeurs aberrantes et d'effets aléatoires non gaussiens, le biais au niveau des estimations des effets aléatoires est plus faible avec le modèle Student/Dirichlet qu'avec le modèle Gauss/Gauss, et ce, d'autant plus que l'amplitude ou la proportion de valeurs aberrantes augmentent. Ainsi la loi de Student permet d'assouplir l'hypothèse de normalité des résidus. Le biais est également plus faible avec le modèle Student/Dirichlet lorsque la distribution des effets aléatoires s'écarte de la normalité. Ainsi, les processus de Dirichlet permettent d'assouplir l'hypothèse de normalité des effets aléatoires.

La précision des estimations est quasiment comparable avec les deux modèles ; les probabilités de couverture des intervalles de crédibilité à 95 % sont proches de 95 % dans les deux cas. Ces résultats incitent à l'utilisation du modèle Student/Dirichlet en présence de valeurs aberrantes et d'effets aléatoires non gaussiens.

Pour les données de PSA après traitement UFHI, les estimations des effets aléatoires obtenues avec les deux modèles étaient relativement différentes pour r_2 et r_4. Les résultats des simulations de l'article incitent à utiliser le modèle Student/Dirichlet pour ces données.

4.2.2.2 Comparaison des performances des marqueurs

Lorsqu'un des effets aléatoires d'un modèle mixte est considéré comme marqueur diagnostique d'une maladie, les résultats des simulations ont montré que les estimations d'AROC sont moins biaisées avec le modèle Student/Dirichlet en présence de valeurs aberrantes et d'effets aléatoires non gaussiens. Ainsi, la réduction du biais au niveau des estimations des marqueurs réduit le biais au niveau de l'estimation de leurs performances diagnostiques. Cette conclusion est valable dans le cas de ces simulations. Un biais au niveau des estimations des marqueurs peut ne pas entraîner de changement dans l'ordre relatif des valeurs des malades et des non malades ; il peut donc ne pas conduire à des changements au niveau des courbes ROC. De même, dans le cas des simulations, le modèle Student/Dirichlet a conduit, dans l'ensemble, à des valeurs d'AROC plus élevées que celles obtenues avec le modèle Gauss/Gauss, mais cette conclusion est spécifique des simulations réalisées. Le modèle Gauss/Gauss conduit dans l'ensemble, par patient, à des estimations un peu plus biaisées de l'effet aléatoire que le modèle Student/Dirichlet. Ceci se traduit, sur l'ensemble des patients, à une erreur de mesure plus élevée dans l'estimation de l'effet aléatoire, ce qui peut expliquer les valeurs d'AROC plus faibles avec le modèle

Gauss/Gauss. Le fait qu'une erreur de mesure dans l'estimation d'un marqueur conduise à une sous estimation de l'AROC est expliqué de manière détaillée dans la partie 4.3.4.

L'AROC obtenue pour la vélocité des PSA à partir des estimations de paramètres du modèle Student/Dirichlet est proche de 0,5. Le nadir de PSA avait une AROC de 0,685, supérieure à celle de la date du nadir (0,630), avec des intervalles de crédibilité quasiment disjoints. Ainsi, le nadir de PSA semble être le meilleur marqueur, parmi la date du nadir et la vélocité, pour discriminer les patients suivant le résultat attendu des biopsies, du moins, le moins mauvais. Les courbes ROC étaient emboîtées pour les 3000 itérations de la chaîne MCMC, ce qui a permis la comparaison directe des marqueurs.

Les AROC obtenues à partir des estimations du modèle Student/Dirichlet et du modèle Gauss/Gauss étaient assez similaires, néanmoins, on peut être plus confiant dans les résultats obtenus à partir du modèle Student/Dirichlet. Un biais dans l'estimation des valeurs d'un marqueur peut entraîner un biais dans l'estimation de ses performances diagnostiques. Le fait de corriger le biais dans l'estimation des valeurs n'a pas pour objectif d'améliorer les performances diagnostiques du marqueur, mais d'obtenir une estimation plus fiable de ses performances, qui soit généralisable à la population dont est issu l'échantillon de l'étude. Les vraies performances du marqueur ne changent pas, l'objectif est uniquement d'obtenir des estimations qui soient les plus proches possibles des vraies valeurs.

4.3 Compléments à l'article

4.3.1 Processus de Dirichlet

Il est mentionné dans l'article que les processus de Dirichlet conduisent à des distributions de probabilités discrètes avec une probabilité de 1 (Muller et Quintana, 2004). A chaque itération, les valeurs d'effets aléatoires r_1 et r_4 sont regroupées par clusters de valeurs identiques. Leur distribution est représentée pour trois itérations de l'algorithme MCMC sur la figure 4.3. Les valeurs indiquées en haut à gauche indiquent le nombres de clusters pour l'itération en question, à comparer au nombre total de patients de 285.

A chaque itération, il existe des zones de valeurs d'effets aléatoires où les clusters sont très nombreux et rapprochés les uns des autres ; dans ces zones, la distribution a posteriori est très proche d'une distribution continue. Néanmoins, il existe d'autres zones, souvent éloignées de la moyenne des valeurs d'effets aléatoires, où les clusters sont bien séparés les uns des autres ; la distribution des valeurs dans ces zones serait difficilement reproductible avec une loi de probabi-

lité continue. Dans l'ensemble, les distributions obtenues s'écartent de la loi normale, en tenant compte parfois d'une sur-dispersion des valeurs ou d'une asymétrie. Dans certains cas, il semble même apparaître des mélanges de distributions. Ceci reflète la grande souplesse des processus de Dirichlet, avec l'inconvénient de fournir des distributions dont le support est constitué d'un nombre fini de points. Il est à noter, tout de même, que le nombre de clusters reste élevé, oscillant entre 50 et 70, permettant ainsi de maintenir une hétérogénéité dans les valeurs d'effets aléatoires estimées.

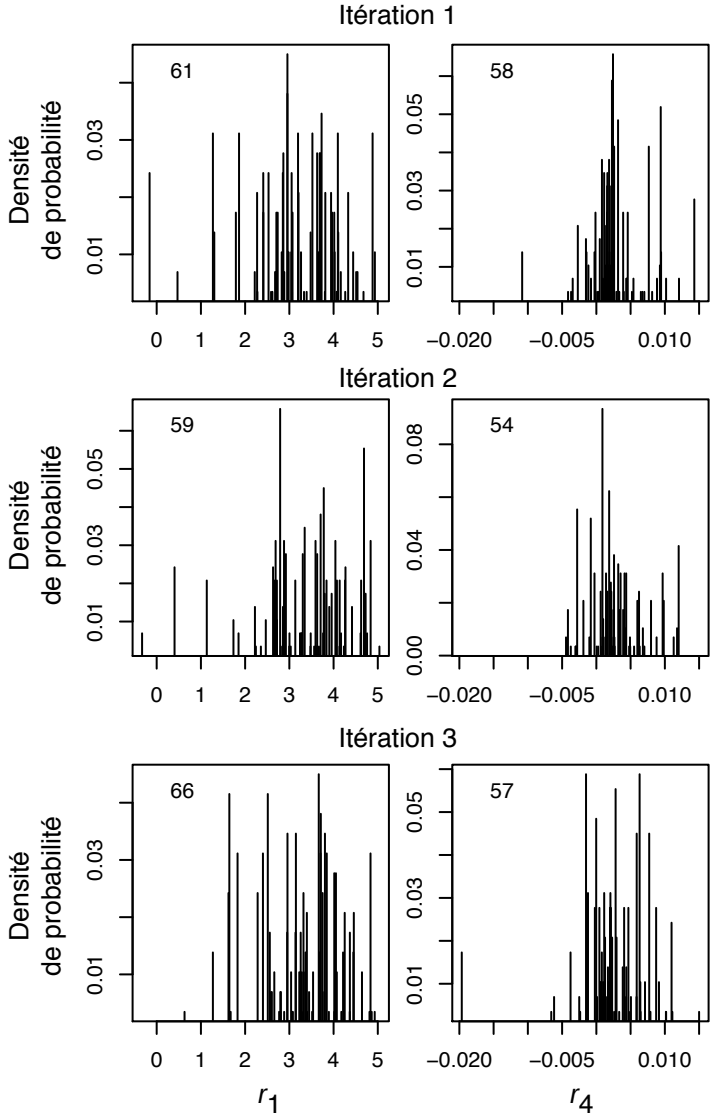

Figure 4.3 – Distributions des valeurs d'effets aléatoires r_1 et r_4 pour trois itérations de l'algorithme MCMC.

4.3.2 Courbes de prédiction

Dans cette étude, l'objectif était de trouver un marqueur permettant de discriminer au mieux les patients dont les biopsies seront positives de ceux dont les biopsies seront négatives. C'est pourquoi les performances des trois marqueurs ont été mesurées au travers de courbes ROC. Néanmoins, pour comparaison, les courbes de prédiction du nadir de PSA et de la date du nadir sont présentées dans la figure 4.4.

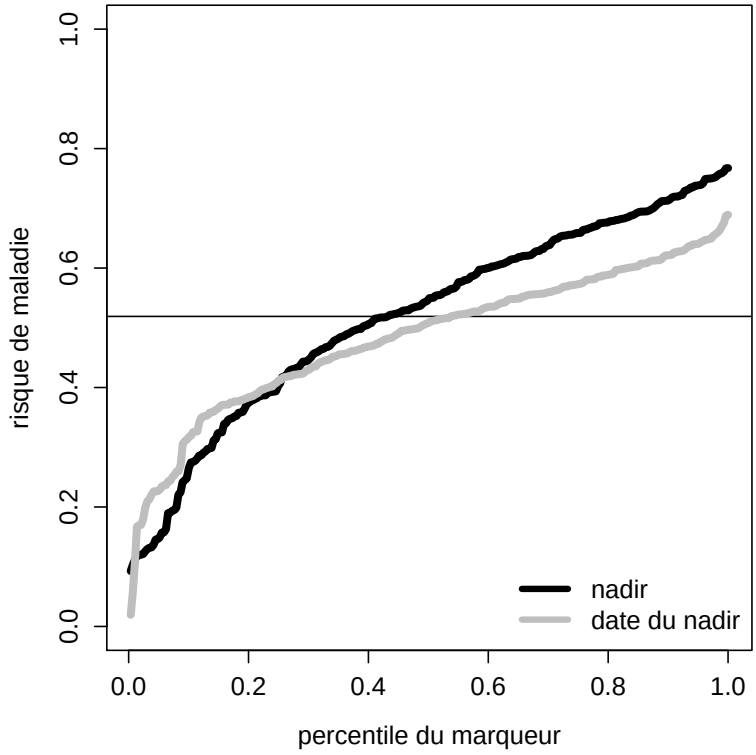

Figure 4.4 – Courbes de prédiction du nadir de PSA et de la date du nadir.

Les risques de maladie en fonction de la valeur du marqueur ont été modélisées à l'aide d'une régression logistique, en transformant la variable explicative – i.e. le marqueur – à l'aide de la méthode de Box-Cox. Le choix de la valeur du paramètre de transformation a été effectué par vraisemblance profilée. Aucun modèle adéquat n'a pu être trouvé pour modéliser le risque de maladie en fonction de la vélocité des PSA, c'est pourquoi la courbe de prédiction associée n'est pas présentée.

Les courbes de prédiction montrent que le nadir de PSA prédit des risques relativement plus élevés que la date du nadir pour une grande partie de la population. Les proportions de risques faibles prédits par les deux marqueurs sont relativement similaires. Les résultats issus de

ces courbes de prédiction semblent favoriser l'utilisation du nadir de PSA pour discriminer les patients suivant le résultat attendu des biopsies.

4.3.3 Intérêt de la modélisation des profils de PSA

Dans l'introduction de cette partie, il a été suggéré que les valeurs de nadir de PSA et de date du nadir calculées directement à partir des mesures de PSA dépendent de la fréquence de ces mesures, ainsi que des augmentations ponctuelles de PSA non liées à la persistance du cancer. Un ensemble de simulations a été réalisé afin de mesurer l'impact de ces deux facteurs sur les estimations des différents marqueurs, ainsi que de comparer les biais obtenus en modélisant les mesures de PSA ou en calculant les paramètres directement à partir des valeurs mesurées de PSA.

4.3.3.1 Plan des simulations

Les profils de PSA ont été simulés pour n patients à partir d'un modèle similaire à celui des données de PSA après traitement UFHI (modèle 4.7) :

$$\ln(y_{ij}+1) = \ln\left(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1\right) + \varepsilon_{ij}, \quad \varepsilon_{ij} \hookrightarrow t\left(\sigma_\varepsilon^2, \nu\right)$$
$$r_{1i} \hookrightarrow \mathcal{N}\left(\mu_{r_1}, \sigma_{r_1}^2\right), \quad r_{2i} \hookrightarrow \mathcal{N}\left(\mu_{r_2}, \sigma_{r_2}^2\right), \quad r_{3i} \hookrightarrow \mathcal{N}\left(\mu_{r_3}, \sigma_{r_3}^2\right), \quad r_{4i} \hookrightarrow \mathcal{N}\left(\mu_{r_4}, \sigma_{r_4}^2\right)$$

Ce modèle introduit la présence de valeurs aberrantes au niveau des résidus, mais par simplicité, les effets aléatoires ont été générés à partir de lois normales. Les mesures ont été réalisées :

- tous les 20 jours durant les 120 premiers jours de suivi (période 1) ;
- tous les 30 jours entre le 120$^{\text{ième}}$ et le 240$^{\text{ième}}$ jours (période 2) ;
- tous les 50 jours par la suite.

Dans la réalité, les mesures étant rarement effectuées le jour prévu, les temps exacts de mesure ont été générés à partir d'une loi uniforme ; la borne gauche de cette loi était la date à laquelle la mesure était censée avoir lieu ; la borne droite était la date théorique de la mesure suivante. La date de la dernière mesure a été générée pour chaque patient à partir d'un échantillon d'une loi de Weibull de paramètre d'échelle 1 et de paramètre de forme 100.

Durant toutes les simulations, les valeurs des effets aléatoires pour les n patients ont été laissées inchangées, mais les valeurs de résidus et les dates de mesures ont été re-générées.

Une fois les profils de PSA simulés, les effets aléatoires, et par là, les valeurs des trois marqueurs ont été estimés pour chacun des patients à l'aide d'un modèle Student/Gauss (loi

de Student pour les résidus, lois normales pour les effets aléatoires), tenant compte ainsi de la présence de valeurs aberrantes au niveau des mesures. Ceci correspond à la méthode dite " par modélisation ". Pour chaque simulation, les valeurs des trois marqueurs ont également été calculées directement à partir des mesures de PSA observées (méthode " directe ") :

- pour le nadir, en retenant la valeur minimale de PSA mesurée ;
- pour la date du nadir, en retenant la date du nadir observé ;
- pour la vélocité, en retenant la pente d'une régression log-linéaire des mesures de PSA effectuées après la date du nadir observé en fonction de la date de la mesure.

Durant les simulations, deux paramètres ont été modifiés :

- le nombre de degrés de liberté de la loi de Student, les valeurs faibles de ν générant plus de valeurs aberrantes ; trois valeurs ont été retenues : $\{1,7\,;\,5\,;\,10\}$, 1,7 correspondant approximativement à la valeur estimée à partir des vrais données de PSA ;
- la fréquence des mesures pour la première période durant laquelle le nadir est atteint par la plupart des patients ; trois fréquences ont été retenues : tous les 10, 15 et 20 jours.

Pour chaque configuration de paramètres, 100 jeux de données ont été simulés pour 100 patients. Le biais relatif moyen lié à l'estimation des trois marqueurs a été calculé pour chaque configuration de paramètres et pour chaque patient, à partir des estimations obtenues par les deux méthodes. Le tableau 4.2 transcrit les quartiles des biais relatifs moyens des 100 patients.

4.3.3.2 Résultats des simulations

Quels que soient le jeu de simulation et le marqueur considérés, les résultats obtenus avec la méthode par modélisation des profils étaient moins biaisés que ceux obtenus à partir de la méthode directe. Le biais de la date du nadir et celui de la vélocité diminuaient pour la méthode directe lorsque la proportion de valeurs aberrantes (fixée par ν) diminuait. Pour la vélocité, ce résultat est en accord avec le fait que l'estimation de la valeur d'une pente peut être biaisée en présence de valeurs aberrantes. En augmentant la fréquence des mesures durant la première période, le biais de la date du nadir estimée par la méthode directe diminuait. Pour le nadir, l'évolution du biais en fonction de la proportion de valeurs aberrantes ou de la fréquence des mesures n'était pas très nette, mais le biais était toujours plus faible avec la méthode par modélisation.

Ces résultats, dans le cas des données de PSA, favorisent la modélisation des profils de biomarqueurs pour estimer les marqueurs reflétant leur cinétique plutôt que le calcul direct des marqueurs à partir des mesures observées. Cette conclusion pourrait être étendue à tous les

marqueurs dont les valeurs dépendent de la fréquence des mesures et de la présence de valeurs aberrantes.

Tableau 4.2 – Résultats des simulations comparant la méthode directe à la méthode par modélisation.

			Date du nadir			Nadir			Vélocité		
ν	fréq	Méthode	25%	50%	75%	25%	50%	75%	25%	50%	75%
1,7	20	Modélisation	-0,030	0,006	0,044	-0,039	-0,019	0,002	-0,322	-0,044	0,125
		Directe	0,486	0,618	0,782	0,851	2,131	4,586	0,089	0,943	2,438
	15	Modélisation	-0,019	0,006	0,035	-0,043	-0,017	0,003	-0,259	-0,072	0,093
		Directe	0,442	0,583	0,759	0,825	2,082	4,545	0,303	1,181	2,572
	10	Modélisation	-0,014	0,007	0,028	-0,031	-0,015	0,002	-0,226	-0,027	0,102
		Directe	0,387	0,572	0,722	0,736	1,961	4,311	0,358	1,497	3,092
5	20	Modélisation	-0,018	0,010	0,031	-0,031	-0,012	0,002	-0,250	-0,044	0,076
		Directe	0,375	0,556	0,711	0,966	2,276	4,937	-0,302	-0,052	0,601
	15	Modélisation	-0,012	0,010	0,023	-0,034	-0,011	0,000	-0,195	-0,030	0,055
		Directe	0,359	0,470	0,632	0,940	2,266	4,851	-0,299	-0,018	0,626
	10	Modélisation	-0,009	0,010	0,021	-0,023	-0,011	-0,001	-0,167	-0,016	0,083
		Directe	0,308	0,450	0,566	0,924	2,208	4,819	-0,306	0,138	0,843
10	20	Modélisation	-0,018	0,010	0,031	-0,032	-0,013	0,002	-0,163	-0,025	0,136
		Directe	0,333	0,480	0,616	0,986	2,325	4,992	-0,391	-0,160	0,263
	15	Modélisation	-0,01	0,010	0,027	-0,028	-0,012	0,000	-0,131	-0,012	0,087
		Directe	0,311	0,435	0,588	0,964	2,311	4,930	-0,387	-0,064	0,418
	10	Modélisation	-0,005	0,009	0,023	-0,020	-0,007	0,003	-0,115	-0,017	0,057
		Directe	0,278	0,409	0,530	0,941	2,265	4,838	-0,416	-0,096	0,527

4.3.3.3 Comparaison des résultats sur les données de PSA

Le tableau 4.3 présente les valeurs d'AROC estimées à partir de la méthode directe et par modélisation. Les résultats obtenus avec la méthode par modélisation sont légèrement différents de ceux présentés dans l'article, car ils sont basés sur 289 patients au lieu de 285 dans l'article.

Tableau 4.3 – AROC estimées par la méthode directe et par modélisation pour les données de PSA.

Critère diagnostique	Méthode	AROC
Nadir	Par modélisation	0,692 [0,685 ; 0,699]
	Directe	0,673 [0,614 ; 0,738]
Date du nadir	Par modélisation	0,635 [0,596 ; 0,670]
	Directe	0,537 [0,470 ; 0,603]
Vélocité	Par modélisation	0,534 [0,486 ; 0,588]
	Directe	0,631 [0,563 ; 0,699]

Les conclusions restent inchangées quant au choix du nadir comme meilleur marqueur, quelle que soit la méthode utilisée, même si la méthode directe semble sous-estimer les performances du nadir. Par contre, pour la date du nadir et la vélocité, les résultats obtenus avec les deux méthodes conduisent à des conclusions totalement différentes. Pour la méthode directe, la date du nadir ne discriminerait pas mieux les patients que le simple hasard, la vélocité serait un marqueur acceptable ; la conclusion opposée est obtenue avec la méthode par modélisation. L'impact de la fréquence des mesures et de la présence de valeurs aberrantes sur le biais des estimations des marqueurs avec la méthode directe est probablement à l'origine de l'inversion des conclusions. On retiendra ici les résultats obtenus par la méthode par modélisation, qui permet de limiter le biais quant aux estimations des valeurs de marqueurs.

4.3.4 Erreurs de mesure et biais des estimations des AROC

Des estimations biaisées des marqueurs peuvent dans certains cas biaiser les estimations des AROC des marqueurs. C'était le cas pour les simulations présentées dans l'article. Cependant, une simple erreur de mesure au niveau du marqueur peut également biaiser les estimations d'AROC.

Considérons pour cela un cas très simple où le marqueur suit une loi normale chez les patients malades ($Y_1 \hookrightarrow \mathcal{N}(\mu_1, \sigma^2)$) et chez les non malades ($Y_0 \hookrightarrow \mathcal{N}(\mu_0, \sigma^2)$), de variance

commune. La méthode de mesure du marqueur n'est pas parfaite ; les mesures réellement effectuées suivent donc des lois normales de variance plus élevée : $Y_1^\star \hookrightarrow \mathcal{N}(\mu_1, \sigma^2 + \delta^2)$ et $Y_0^\star \hookrightarrow \mathcal{N}(\mu_0, \sigma^2 + \delta^2)$, où δ^2 caractérise l'erreur de mesure.

L'AROC, pour les marqueurs sans erreur de mesure, est donnée par :

$$AROC = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{2}\sigma}\right)$$

où $\Phi(.)$ correspond à la fonction de répartition de la loi normale. Pour les valeurs de marqueurs mesurées avec erreurs, l'AROC est :

$$AROC^\star = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{2}\sqrt{\sigma^2 + \delta^2}}\right)$$

Le rapport entre les aires avec et sans erreur de mesure n'est pas calculable analytiquement, mais peut être représenté graphiquement pour un ensemble de valeurs de σ^2 et de δ^2. La figure 4.5 présente le rapport $AROC/AROC^\star$ pour $\mu_1 = 1$ et $\mu_0 = 0$. Ce ratio est toujours supérieur à 1, indiquant que l'AROC obtenue avec erreur de mesure est toujours inférieure à celle obtenue sans erreur de mesure, des résultats similaires étant obtenus lorsque la variance de l'erreur de mesure est proche de zéro, ou négligeable devant la variabilité propre du marqueur chez les malades et les non malades.

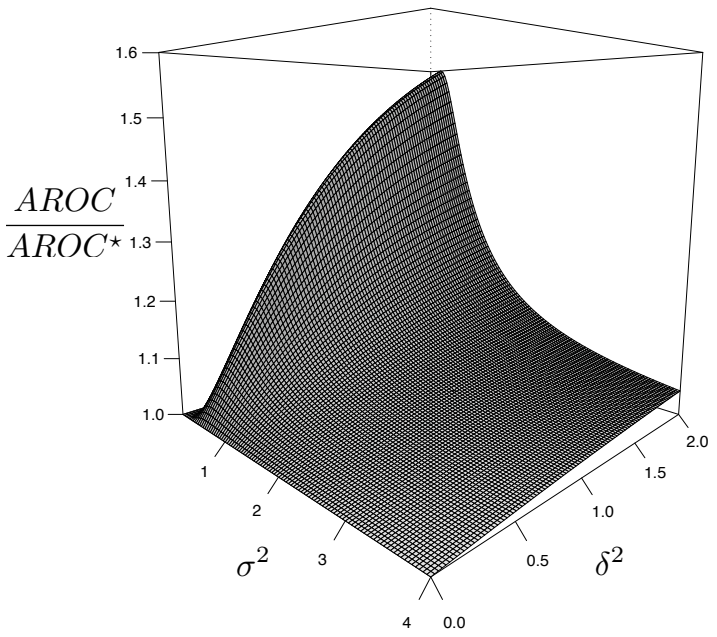

Figure 4.5 – Rapport des AROC sans erreur de mesure et avec erreur de mesure.

Dans le cas plus général où aucune distribution spécifique n'est fixée pour les marqueurs, mais où une erreur gaussienne et de même ampleur est ajoutée aux mesures dans les deux groupes, Coffin et Sukhatme (1997) ont montré, grâce à des développements de Taylor, que les AROC avec erreur de mesure sont plus faibles que celles sans erreur de mesure, sauf lorsque la distribution du marqueur est identique chez les malades et les non malades.

Dans le cas de l'étude sur les PSA après UFHI, les valeurs de marqueurs sont estimées à partir des profils modélisés de PSA. Le modèle Student/Dirichlet permet de limiter le biais au niveau des estimations des marqueurs, mais il reste tout de même l'imprécision liée à l'estimation des paramètres. Ceci explique notamment pourquoi le biais relatif des estimations d'AROC obtenues lors des simulations avec le modèle Student/Dirichlet reste non négligeable, bien que plus faible que celui obtenu avec le modèle Gauss/Gauss.

Dans l'article, l'effet de l'imprécision des mesures des marqueurs a été retranscrit dans les intervalles de crédibilité des AROC. Dans le cas des nadirs de PSA, soit n_1 le nombre de patients malades, n_0 le nombre de patients non malades et $\mathbf{y}$ les mesures de PSA sur l'ensemble des patients. Si les valeurs de nadir chez les malades et non malades, notées respectivement $\mathbf{nadir}_1$ et $\mathbf{nadir}_0$, étaient connues sans erreur de mesure, une estimation de l'AROC serait obtenue par :

$$AROC = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(nadir_{1i} > nadir_{0j}) + 1/2 \times I(nadir_{1i} = nadir_{0j})$$

Les valeurs de nadir par patient n'étaient pas connues de manière exacte, mais calculées à partir de leur distribution a posteriori $P(\mathbf{nadir}_1, \mathbf{nadir}_0|\mathbf{y})$. Ainsi :

$$AROC|\mathbf{y} = \int\int \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} (I(nadir_{1i} > nadir_{0j}) + 1/2 \times I(nadir_{1i} = nadir_{0j})) \times P(\mathbf{nadir}_1, \mathbf{nadir}_0|\mathbf{y})\, d\mathbf{nadir}_1 d\mathbf{nadir}_0$$

La distribution a posteriori des nadirs sachant les mesures de PSA a été approximée par les échantillons de 3000 valeurs de nadirs par patient obtenues par MCMC, conduisant à un échantillon de 3000 valeurs d'AROC issues de la distribution a posteriori de celle-ci sachant les données. La distribution a posteriori de l'AROC intègre l'incertitude sur l'estimation des valeurs de nadir au travers de la distribution a posteriori des valeurs de nadir sachant les mesures de PSA. Ainsi, l'intervalle de crédibilité de l'AROC obtenu, soit par la méthode des quantiles, soit par la méthode HDP, intégrait l'incertitude quant à l'estimation des valeurs de nadirs.

4.4 Bilan du chapitre 4

Dans cette partie il a été montré que, lorsque l'objectif est de comparer les performances diagnostiques ou pronostiques de marqueurs reflétant la cinétique d'un biomarqueur, il est conseillé de calculer les valeurs de marqueurs à partir de la modélisation des profils du biomarqueur. Ceci est d'autant plus vrai lorsque les valeurs de marqueurs sont sensibles à la fréquence des mesures et à la présence de valeurs aberrantes.

Une méthode robuste a été proposée pour modéliser les profils de biomarqueurs longitudinaux, combinant la loi de Student et les processus de Dirichlet. Cette méthode n'est pas spécifique aux données de PSA. Dans de nombreux cas, lorsque des biomarqueurs longitudinaux sont modélisés grâce à un modèle à effets mixtes, les effets aléatoires ne suivent pas des distributions gaussiennes, d'où l'intérêt des processus de Dirichlet. La présence de valeurs aberrantes aussi marquées que celles observés dans les données de PSA n'est peut être pas fréquente, mais il est rare qu'il n'y ait pas d'erreurs de mesure, d'où l'intérêt de la loi de Student.

D'autres lois auraient pu être utilisées pour assouplir l'hypothèse de normalité des résidus, comme la loi normale contaminée ou la distribution dite " slash distribution " (Rosa et *al.*, 2003). Le choix de la loi dépend en partie de la proportion de valeurs aberrantes et de leur amplitude, mais la loi de Student reste toujours une bonne solution par rapport à la loi normale et est relativement facile à implémenter.

De même, les processus de Dirichlet ne constituent pas la seule solution pour assouplir l'hypothèse de normalité des effets aléatoires. Les lois normales et de Student asymétriques auraient pu être envisagées (Lee et Thompson, 2008), mais elles contraignent plus la distribution des effets aléatoires que les processus de Dirichlet. Une critique souvent formulée à l'égard de ces processus est qu'ils conduisent à une distribution de probabilité discrète avec une probabilité 1 (Muller et Quintana, 2004). Des méthodes non paramétriques ou semi-paramétriques, mais qui conduisent à des distributions de probabilité continues, sont les mélanges de processus de Dirichlet (MacEachern et Müller, 1998 ; Ohlssen et *al.*, 2007) ou les arbres Polya (Lavine, 1992, 1994). Néanmoins, le fait que les processus de Dirichlet conduisent à des distributions de probabilité discrètes ne modifie pas forcément l'ordre relatif des valeurs de marqueurs des malades et des non malades ; or c'est uniquement cet ordre qui est utilisé pour la construction des courbes ROC et non les valeurs réelles des marqueurs. De plus, les arbres Polya sont très demandeurs d'un point de vue calculatoire ; c'est pourquoi les processus de Dirichlet ont été privilégiés.

Il est rare qu'un modèle mixte assouplisse à la fois la distribution des résidus et celle des effets aléatoires. A notre connaissance, ce cas n'avait été considéré auparavant que par Pinheiro et *al.* (2001) qui avaient utilisé une distribution de Student à la fois pour les résidus et les effets aléatoires. Dans le cadre des données de PSA, à l'exception de Proust-Lima et Taylor (2009) qui avaient considéré un mélange de lois normales afin de caractériser la distribution des effets aléatoires, c'est la première fois qu'autant d'attention est apportée à la modélisation de ce type de données. Le travail réalisé a montré l'impact que ceci peut avoir au niveau des conclusions. Les résultats obtenus avec le modèle Student/Dirichlet privilégient l'utilisation du nadir de PSA pour discriminer les patients selon le résultat attendu des biopsies.

Le statut des patients n'a pas été introduit dans la modélisation des profils du biomarqueur. En réalité, il est difficile de savoir exactement sur quels paramètres le statut agit, même s'il semble avoir un effet sur le nadir et la date du nadir. Plutôt que de rajouter des contraintes supplémentaires au modèle, en n'étant pas sûr de l'endroit où il faut les rajouter, il a été préféré d'utiliser des modèles souples – avec l'utilisation de processus de Dirichlet – pouvant reproduire n'importe quel type de profil, indépendamment du statut du patient. L'hypothèse est que le modèle est suffisamment souple ; ainsi, les résultats obtenus en termes de comparaison des performances des marqueurs sont les mêmes que ceux qui auraient pu être obtenus en introduisant correctement l'effet du statut des patients dans le modèle. Le type d'analyse proposé est uniquement descriptif et ne permet pas, contrairement aux modèles biologiques, de mieux comprendre l'effet de la maladie sur les PSA.

Troisième partie

Définition de seuils de marqueurs

Chapitre

5 Estimation du seuil optimal et de son intervalle de confiance

La partie précédente a montré que le nadir est le meilleur marqueur issu de la cinétique des PSA, parmi la date du nadir et la vélocité, pour discriminer les patients selon le résultat attendu des biopsies (il faut tout de même garder à l'esprit que ses performances diagnostiques restent limitées). Les courbes ROC des différents marqueurs étaient emboîtées. Ainsi, quel que soit le niveau de spécificité considéré, le nadir de PSA a une meilleure sensibilité que les autres marqueurs. Pour tout seuil de nadir de PSA retenu, l'utilité espérée en utilisant ce test pour prendre une décision est meilleure que celle des autres marqueurs, quel que soit le seuil retenu pour ces derniers. Néanmoins, il reste à définir une valeur seuil du nadir de PSA au dessus de laquelle les cliniciens ont intérêt à réaliser une biopsie. Ceci a fait l'objet du second article de thèse et de travaux complémentaires présentés dans cette partie.

5.1 Utilité espérée : entre risque et incertitude

5.1.1 Utilité espérée pour le choix du seuil optimal

Il a été montré que le seuil optimal d'un marqueur est le seuil c qui maximise l'utilité espérée du test lorsque le marqueur est utilisé dans une population (partie 2.3.1) :

$$\mathcal{U}(c) = \text{Sen}(c)\pi u(z_{TM}) + (1-\text{Sen}(c))\pi u(z_{\bar{T}M}) + (1-\text{Spe}(c))(1-\pi)u(z_{T\bar{M}}) + \text{Spe}(c)(1-\pi)u(z_{\bar{T}\bar{M}})$$

Ici, la distribution de probabilité conjointe des valeurs de marqueurs et du statut des patients a été décomposée selon l'approche orientée marqueur, faisant intervenir la sensibilité, la spécificité et la prévalence. Cette factorisation est fréquemment retenue pour l'estimation du seuil optimal d'un marqueur, en raison de la facilité à calculer la sensibilité, la spécificité et la prévalence.

Le seuil qui maximise la fonction précédente est également le seuil qui maximise la fonction :

$$\mathcal{U}^\star(c) = \text{Sen}(c) + \text{Spe}(c)\frac{CN}{BN}\frac{1-\pi}{\pi}$$

où BN correspond au bénéfice net de traiter un patient malade par rapport à ne pas le traiter ($BN = u(z_{TM}) - u(z_{\bar{T}M})$) et CN correspond au coût net de traiter un patient non malade par rapport à ne pas le traiter ($CN = u(z_{T\bar{M}}) - u(z_{\bar{T}\bar{M}})$). Il est rappelé que ces ratios et coûts peuvent être estimés pour un individu, ou pour un ensemble de patients ; dans le premier cas, le seuil déterminé est optimal pour un patient, alors que dans le second, il est optimal en moyenne sur un ensemble de patients. Ce seuil optimal $c^\star$ est tel que la dérivée de $U^\star$ par rapport à c est nulle en ce point. C'est donc la valeur c telle que :

$$\frac{d\text{Sen}(c)/dc}{d\text{Spe}(c)/dc} = -\frac{CN}{BN}\frac{1-\pi}{\pi} \quad \Leftrightarrow \quad \frac{f_1(c)}{f_0(c)} = \frac{CN}{BN}\frac{1-\pi}{\pi} \tag{5.1}$$

en s'assurant que la valeur obtenue est bien le maximum global de la fonction d'utilité. Par la suite, on notera $R = CN/BN \times (1-\pi)/\pi$.

Lorsque la prévalence est de 0,5 et que le ratio bénéfice net sur coût net est de 1, le seuil optimal est la valeur du marqueur telle que les densités de probabilité des valeurs du marqueur chez les malades et les non malades soient égales. Si les densités de probabilité sont représentées sur un même graphique, le seuil optimal est le point d'intersection des deux courbes (figure 5.1, traits légers).

Dans le cas où la prévalence est de 1/4, le seuil optimal est la valeur du marqueur telle que le rapport entre les deux densités de probabilité soit de $(3/4)/(1/4) = 3$. C'est également le point d'intersection entre la courbe associée à f_0 multipliée par 3/4 et la courbe associée à f_1 multipliée par 1/4 ; ces courbes ne correspondent plus à des densités de probabilité (figure 5.1, traits gras). Par rapport à la situation où la prévalence était de 0,5, le seuil optimal est plus élevé. La population est constituée de plus de non malades que de malades. L'utilité espérée étant calculée sur l'ensemble de la population, elle est plus influencée, dans ce cas, par les résultats obtenus chez les non malades, pour lesquels il vaut mieux un seuil plus élevé afin qu'ils ne soient pas traités. Cet exemple souligne le fait que le seuil obtenu est un seuil qui maximise l'utilité

moyenne sur une population ; c'est un seuil de décision pour une population et non pas un seuil pour un patient. Néanmoins, le ratio BN/CN peut être déterminé en fonction des préférences d'un unique patient ; dans ce cas, le seuil est optimal pour l'ensemble des patients exprimant les mêmes préférences en termes de qualité ou de quantité de vie.

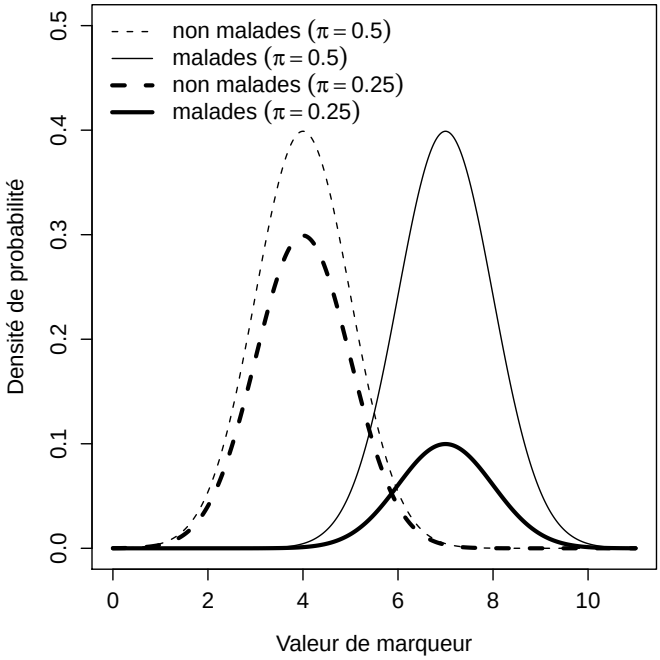

Figure 5.1 – Densités de probabilité des valeurs de marqueurs chez les malades et les non malades pour un ratio bénéfice net sur coût net de 1 ; les courbes en traits légers correspondent au cas $\pi = 0,5$, celles en traits gras au cas $\pi = 0,25$.

La représentation graphique de la figure 5.1 permet de visualiser facilement l'évolution du seuil en fonction de la prévalence. Cette même représentation est utilisable pour tenir compte du ratio bénéfice net sur coût net. Supposons que la prévalence soit de 0,5 et que le ratio bénéfice net sur coût net soit maintenant de 2. Le seuil optimal est la valeur de marqueur telle que le rapport entre les deux densités de probabilité soit de 2. C'est également le point d'intersection entre la courbe associée à f_0 et celle associée à f_1 multipliée par 2 (figure 5.2, traits gras). Le seuil obtenu est plus petit que pour un ratio bénéfice net sur coût net de 1. Avec un ratio de 2, il y a un grand bénéfice à traiter les patients malades par rapport à ne pas les traiter, le coût des patients traités à tort étant négligeable ; ainsi, il faut privilégier un seuil faible, pour que le maximum de patients malades soient détectés. Lorsque les courbes obtenues se croisent en plus d'un seul point – cas où la fonction d'utilité présente plusieurs maxima – le seuil optimal est celui qui conduit à la plus grande utilité espérée.

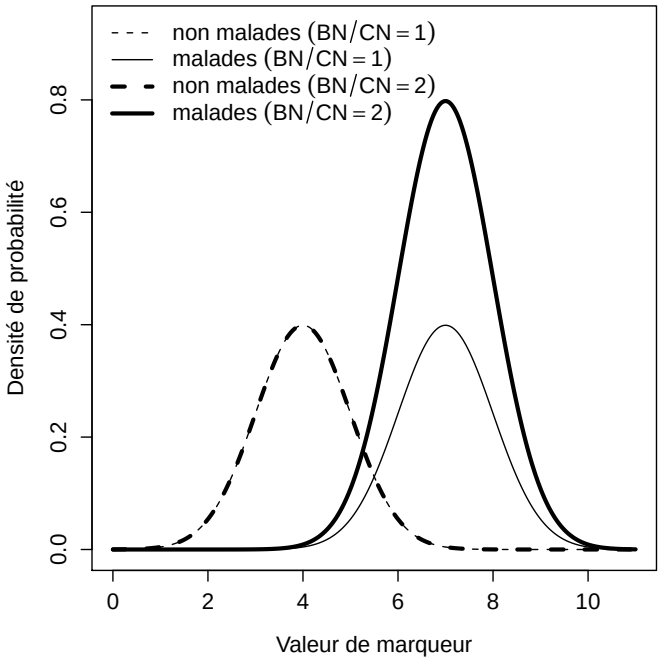

Figure 5.2 – Densités de probabilité des valeurs de marqueurs chez les malades et les non malades pour une prévalence de 0,5 ; les courbes en traits légers correspondent au cas $BN/CN = 1$, celles en traits gras au cas $BN/CN = 2$.

Le seuil optimal est également, d'après l'équation (5.1), la valeur du marqueur telle que la tangente à la courbe ROC en ce point ait une pente égale à $(CN/BN)\times(1-\pi)/\pi$. Contrairement à ce qui peut être lu dans de nombreux articles, ce n'est donc pas le point de la courbe ROC le plus proche du point de coordonnées (0,1). Le seuil associé à ce dernier point est en réalité le seuil qui maximise l'indice de Youden, défini par :

$$\text{Youden}(c) = \text{Sen}(c) + \text{Spe}(c) - 1$$

Cette dernière approche est valable lorsque la prévalence est de 0,5 et que le ratio bénéfice net sur coût net est de 1, ou bien plus généralement lorsque $(CN/BN) \times (1 - \pi)/\pi = 1$. Dans le cas contraire, le seuil qui maximise l'indice de Youden ne tient pas compte de la prévalence ni des utilités associées aux différentes situations ; il donne juste la valeur de seuil qui sépare au mieux les courbes de densité de probabilité de valeurs du marqueur chez les malades et les non malades.

On considère le cas où le marqueur suit une loi normale de moyenne 0 chez les non malades et 1 chez les malades, l'écart type étant de 0,5 dans les deux groupes. La valeur de seuil maximisant l'indice de Youden est 0,5. Si la prévalence dans la population étudiée est de 0,5,

sur un ensemble de 1000 personnes, en moyenne, 500 ont une valeur de marqueur supérieure à 0,5 et sont traitées ; parmi ces personnes, 79 le sont à tort. Ce nombre de personnes traitées à tort dépend de la prévalence. En effet, pour une prévalence plus faible, de 0,1, uniquement 227 personnes sont traitées, mais parmi elles, 143 le sont à tort. Si l'estimation du seuil avait tenu compte de la prévalence, ce dernier aurait été de 1,05, conduisant à traiter 62 patients ; parmi eux, uniquement 46 l'auraient été à tort, contre 143 dans le cas précédent, mais au risque qu'un plus grand nombre de personnes réellement malades ne soient pas soignées. Dans ce dernier cas, le nombre de malades non traités serait de 54, contre 16 en utilisant le seuil de 0,5. Si le bénéfice net de traiter un patient malade est jugé de 1 et le coût net de 2 (sur une échelle arbitraire), il est préférable d'avoir moins de patients traités à tort que de patient non traités alors qu'ils auraient dû l'être. L'indice de Youden ne tient en aucun cas compte de ces caractéristiques. Le choix du seuil optimal dépend donc de la prévalence et des coûts et bénéfices associés au traitement. L'utilisation de l'indice de Youden pour déterminer ce seuil n'est une approche rationnelle que lorsque $(CN/BN) \times (1-\pi)/\pi = 1$, situation qui n'est pas très fréquente.

Les graphiques 5.1 et 5.2 semblent particulièrement adaptés pour visualiser l'effet de la prévalence et du ratio bénéfice net sur coût net que l'évolution de la tangente à la courbe ROC.

5.1.2 Cas de distributions gaussiennes du marqueur

On suppose que le marqueur suit des lois normales chez les non malades et les malades, de paramètres $\mathcal{N}(\mu_0, \sigma_0^2)$ et $\mathcal{N}(\mu_1, \sigma_1^2)$. Dans ce cas :

$$\begin{aligned} U(c)^\star &= \text{Sen}(c) + \text{Spe}(c) \times R \\ &= P(Y > c | M = 1) + P(Y \leq c | M = 0) \times R \\ &= \left(1 - \Phi\left(\frac{c-\mu_1}{\sigma_1}\right)\right) + \Phi\left(\frac{c-\mu_0}{\sigma_0}\right) \times R \end{aligned}$$

Une formule explicite de la valeur du seuil qui annule la fonction d'utilité existe (Schisterman et Perkins, 2007) :

$$c = \frac{\mu_0(b^2-1) - a + b\sqrt{a^2 + (b^2-1)\sigma_0^2 \ln(b^2 R^2)}}{b^2 - 1} \tag{5.2}$$

avec $a = \mu_1 - \mu_0$ et $b = \sigma_1/\sigma_0$. En réalité, il existe une seconde valeur qui annule la dérivée de la fonction d'utilité, mais elle est sous optimale lorsque les valeurs de marqueurs sont plus élevées chez les malades que chez les non malades. Une estimation ponctuelle du seuil peut être obtenue en remplaçant les paramètres dans la formule (5.2) par leur estimation du maximum

de vraisemblance :

$$\hat{c}^\star = \frac{\hat{\mu}_0(\hat{b}^2-1) - \hat{a} + \hat{b}\sqrt{\hat{a}^2 + (\hat{b}^2-1)\hat{\sigma}_0^2 \ln(\hat{b}^2 R^2)}}{\hat{b}^2 - 1}$$

Un intervalle de confiance est constructible par la méthode Delta (Casella et Berger, 2002) ; elle consiste à supposer que l'estimateur du seuil optimal suit une loi normale et à approximer la variance de cet estimateur par un développement de Taylor d'ordre un. Ainsi :

$$\mathrm{Var}(\hat{c}^\star) = \left(\frac{\partial \hat{c}}{\partial \mu_1}\right)^2 \mathrm{Var}(\mu_1) + \left(\frac{\partial \hat{c}}{\partial \mu_0}\right)^2 \mathrm{Var}(\mu_0) + \left(\frac{\partial \hat{c}}{\partial \sigma_1}\right)^2 \mathrm{Var}(\sigma_1) + \left(\frac{\partial \hat{c}}{\partial \sigma_0}\right)^2 \mathrm{Var}(\sigma_0)$$

Cette méthode peut être étendue aux cas de lois log normales (Leeflang et *al.*, 2008), de lois normales après transformation de Box-Cox (Fluss et *al.*, 2005 ; Schisterman et *al.*, 2008) et de lois gamma, mais, dans ce dernier cas, uniquement pour certaines plages de valeurs de paramètres. Elle est appelée la méthode du plug-in. L'inconvénient est que l'approximation normale de la distribution de l'estimateur du seuil optimal n'est valable que de manière asymptotique, or les études sur les marqueurs diagnostiques ou pronostiques sont souvent réalisées sur des échantillons de taille limitée.

5.1.3 Cas général

Pour certaines lois de probabilité, il n'existe pas forcément de formule explicite du seuil optimal. Dans ce cas, le seuil optimal peut être obtenu par maximisation numérique de la fonction d'utilité, par exemple par un algorithme de type Newton-Raphson. Il n'y a alors plus de formule explicite du seuil optimal ; la seule solution proposée actuellement pour construire un intervalle de confiance consiste à utiliser la méthode du bootstrap. La probabilité de couverture de l'intervalle de confiance ainsi construit n'est pas toujours acceptable (Schisterman et Perkins, 2007), d'où la nécessité de développer de nouvelles méthodes.

5.1.4 Incertitude et seuil optimal

5.1.4.1 Le risque et l'incertitude

Dans la partie précédente, l'estimation ponctuelle du seuil optimal est effectuée comme si les paramètres de distribution des marqueurs étaient connus de manière exacte, alors qu'ils résultent d'une estimation sur un échantillon de population. L'incertitude sur les estimations des paramètres n'intervient que dans la construction de l'intervalle de confiance, au travers des

variances des estimateurs. L'estimation ponctuelle tient compte uniquement de la variabilité entre individus, retranscrite par la sensibilité et la spécificité. Cette variabilité entraîne que tous les malades – ou les non malades – n'ont pas la même valeur de marqueur. Comme indiqué dans l'introduction sur la théorie de la décision (partie 2.2.5), c'est cette variabilité qui est à l'origine de la prise de décision en situation de *risque*. Sans cette variabilité, il n'y aurait aucun risque à prendre une décision.

Deux sources de variabilité ont donc été recensées :

- le *risque*, lié aux variations des caractéristiques des patients au sein des deux groupes de la population, se traduisant par des variations des valeurs de marqueur ;
- et l'*incertitude*, qui est due au fait que les paramètres de distribution des marqueurs dans les deux groupes, ainsi que la prévalence, ne sont pas connus de manière exacte, mais estimés à partir d'un échantillon de la population.

Bien que ces deux sources de variabilité aient des causes très différentes, elles peuvent être caractérisées toutes deux par des distributions de probabilité. Le risque, noté $P_a(z|\boldsymbol{\theta})$, est caractérisé par la sensibilité, la spécificité et la prévalence, $\boldsymbol{\theta}$ dénotant l'ensemble de ces trois paramètres. L'incertitude peut être caractérisée par la distribution a posteriori des paramètres de distribution du marqueur dans les deux groupes et de la prévalence ; elle est notée ici $P(\boldsymbol{\theta}|\mathbf{y})$. D'après le théorème de Bayes, cette distribution a posteriori des paramètres est reliée à la vraisemblance des données et à la probabilité a priori des paramètres : $P(\boldsymbol{\theta}|\mathbf{y}) \propto P(\mathbf{y}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})$. L'information a priori sur les paramètres peut éventuellement être non informative.

L'estimation ponctuelle du seuil optimal présentée dans la partie précédente tient compte du risque, mais pas de l'incertitude. La théorie de la décision Bayésienne permet de réconcilier ces deux aspects.

5.1.4.2 Théorie de la décision Bayésienne

La théorie de la décision Bayésienne vise à prendre des décisions optimales en tenant compte de toutes les formes de variabilité dans le calcul de l'utilité espérée, c'est à dire du risque et de l'incertitude. Cette théorie remonte à Savage (Savage, 1954) ; une bonne introduction en est faite par Dorfman (Dorfman, 1997) et Berger (Berger, 1980). L'utilité espérée ne tient plus compte uniquement du risque ($P_a(z|\boldsymbol{\theta})$) et des utilités ($u(z)$) comme dans l'équation (2.3), mais également de l'incertitude ($P(\boldsymbol{\theta}|\mathbf{y})$), en intégrant cette incertitude sur l'ensemble des valeurs

possibles des paramètres :

$$\mathcal{U}(a) = \int_{\mathcal{Z}} \int_{\Theta} u(z) P_a(z|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}) \, dz d\boldsymbol{\theta} \tag{5.3}$$

Ceci rend les calculs d'utilité espérée nettement plus complexes, mais des solutions seront envisagées via les méthodes MCMC.

Selon Dorfman (1997), le " principe d'équivalence des certitudes " stipule que le fait de ne pas tenir compte de l'incertitude dans le calcul de l'utilité espérée conduit à prendre des décisions similaires à celles obtenues en en tenant compte si et seulement si la distribution a posteriori des paramètres est normale, ainsi que si la fonction à optimiser est linéaire-quadratique selon les paramètres incertains. La seconde condition est rarement valable dans le cas de l'estimation du seuil optimal d'un marqueur. Ainsi, l'incertitude liée à l'estimation des paramètres n'a pas un impact uniquement sur l'intervalle de confiance du seuil optimal, mais également sur l'estimation ponctuelle de ce seuil. Selon Bradlaw et *al.* (2004), le fait de ne pas tenir compte correctement de l'incertitude sur les paramètres entraîne une surestimation de la quantité d'information contenue dans les données et, en général, une surévaluation des utilités. Néanmoins, lorsque la quantité de données augmente, les écarts entre les deux méthodes deviennent minimes (Berger, 1980). Par la suite, la méthode intégrant l'incertitude quant à l'estimation des paramètres de distribution sera appelée la méthode *prédictive* ; la première méthode sera appelée la méthode *plug-in*.

Considérons le calcul de la sensibilité dans le cas où le marqueur suit une loi normale chez les malades et où un a priori non informatif est retenu pour les paramètres de la distribution ($P(\mu_1, \sigma_1^2) = 1/\sigma_1^2$). Sachant les valeurs de marqueurs $\mathbf{y}_1$ mesurées dans l'échantillon de patients malades, la distribution prédite pour la valeur de marqueur d'un nouveau malade, $\tilde{y}_1$, est donnée par :

$$P(\tilde{y}_1|\mathbf{y}_1) \propto \int_{\Theta} P(\tilde{y}_1|\boldsymbol{\theta}_1) \times P(\boldsymbol{\theta}_1|\mathbf{y}_1) \, d\boldsymbol{\theta}$$

où $\boldsymbol{\theta}_1 = \{\mu_1, \sigma_1^2\}$. Dans ce cas précis, la distribution prédite est une distribution de Student :

$$P(\tilde{y}_1|\mathbf{y}_1) \hookrightarrow t\left(\frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j}, \frac{n_1 \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2}{(n_1+1)(n_1-1)}\right)$$

Ainsi, la sensibilité n'est plus obtenue par la fonction de répartition d'une loi normale, mais par celle d'une loi de Student. Le même résultat est obtenu pour la spécificité. Le seuil optimal, dans le cas où les nombres de malades et de non malades sont égaux, où la prévalence est de 0,5 et

où le ratio bénéfice net sur coût net est de 1 est donné par :

$$c^\star = \frac{\mu_0(\nu b^2 - 1) - a + b\sqrt{\nu a^2 + (1-\nu)(n-1)(n+1)\sigma_0^2(\nu b^2 - 1)/n}}{\nu b^2 - 1}$$

avec $\nu = (\sigma_0/\sigma_1)^{2/(n-1)}$. Cette expression est différente de celle obtenue avec la méthode plug-in (équation (5.2)). Quand le nombre de patients par groupe tend vers l'infini, la loi de Student tend vers la loi normale ; les écarts entre le seuil optimal obtenu avec la méthode prédictive et la méthode plug-in deviennent minimes. Un exemple est donné dans la figure 5.3, en prenant $\mu_0 = 0, \mu_1 = 1, \sigma_0 = 0,1$ et $\sigma_1 = 0,2$.

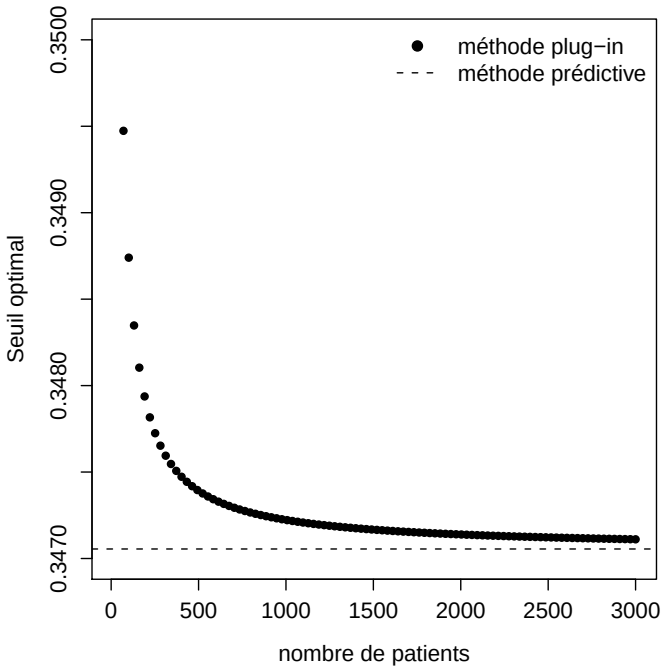

Figure 5.3 – Seuil optimal en fonction du nombre de patients par groupe obtenu avec la méthode prédictive et la méthode plug-in.

Cet exemple montre l'intérêt de tenir compte, au moins d'un point de vue conceptuel, de l'incertitude sur les estimations des paramètres dans l'estimation ponctuelle du seuil optimal. Pour l'intervalle de confiance, la méthode Delta pourrait à nouveau être utilisée.

Une des principales limites de l'approche précédente est que, dans certains cas, l'information a priori sur les paramètres n'est pas conjuguée avec la vraisemblance. Il n'y a donc plus d'expression explicite de la distribution prédictive a posteriori des valeurs de marqueurs ; d'autres solutions que les solutions analytiques sont à envisager. L'intérêt des méthodes MCMC dans ce type de problème est exploré par la suite, avec deux approches différentes.

5.2 Estimation du seuil optimal pour un marqueur fixe

Le problème de l'estimation du seuil optimal d'un marqueur fixe en utilisant l'approche prédictive est un problème d'optimisation de fonction complexe, faisant intervenir des intégrales. De manière générale, la fonction à optimiser peut être notée :

$$\mathcal{U}^{\star}(c) = \mathrm{E}(V(c, \boldsymbol{\theta})) = \int_{\Theta} V(c, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \tag{5.4}$$

où V est l'utilité lorsque les paramètres $\boldsymbol{\theta}$ sont connus. Le calcul de cette intégrale peut être effectué à l'aide de la méthode de Monte Carlo (partie A.4), en échantillonnant des valeurs de $\boldsymbol{\theta}$ dans leur distribution a posteriori $P(\boldsymbol{\theta}|\mathbf{y})$, puis en approximant l'intégrale par la moyenne des fonctions d'utilité calculées pour les différents échantillons de paramètres.

Deux approches sont envisageables ensuite pour estimer le seuil optimal :

- retenir le seuil qui maximise la moyenne des fonctions d'utilité (maximisation de l'utilité espérée moyenne) ;
- retenir la moyenne des seuils optimaux obtenus pour chacune des fonctions d'utilité (moyenne des maxima des fonctions d'utilité).

Ces deux approches sont détaillées par la suite.

5.2.1 Maximisation de l'utilité espérée moyenne

Le principe de cette méthode est de retenir comme seuil optimal le seuil qui maximise la moyenne des fonctions d'utilité sur les différents échantillons issus de la distribution a posteriori des paramètres de distribution des marqueurs. Reste à montrer que ce seuil estimé converge vers le vrai seuil, puis à obtenir un intervalle de confiance de cette estimation ponctuelle.

5.2.1.1 Convergence en probabilité du maximum de la fonction d'utilité moyenne

Montrer que le seuil qui maximise la moyenne des fonctions d'utilité converge vers le vrai seuil nécessite de faire un certain nombre d'hypothèses sur la chaîne MCMC permettant l'échantillonnage des valeurs de paramètres dans leur distribution a posteriori. Cette partie est assez technique. L'objectif n'est pas d'expliciter de façon formelle les propriétés des chaînes MCMC, mais plutôt d'en donner une idée relativement intuitive. Une approche plus formelle peut être trouvée dans Gilks et *al.* (1996) et Robert et Casella (1999).

On considère pour le paramètre θ une distribution de probabilité ϕ définie sur le domaine Θ. Une chaîne MCMC est dite *irréductible* si elle est capable d'atteindre n'importe quel point de Θ pour lequel $\phi(\theta) > 0$. Une chaîne est *périodique* si certaines portions de l'espace ne peuvent être visitées qu'à intervalle de temps régulier ; dans le cas contraire, la chaîne est dite *apériodique*. Une chaîne irréductible peut en principe atteindre n'importe quel point de l'espace pour lequel $\phi(\theta) > 0$, mais cela ne veut pas dire qu'elle l'atteindra forcément. Si la probabilité que la chaîne revienne infiniment souvent en tous points du domaine tels que $\phi(\theta) > 0$ est supérieure à 0 – et vaut 1 à l'exception de certaines valeurs de θ dépendant de ϕ – alors la chaîne est dite *récurrente*. Une distribution $\phi^\star$ est dite stationnaire pour la chaîne si, lorsque la chaîne part d'un point où les valeurs de θ sont distribuées selon $\phi^\star$, alors les valeurs de θ parcourues au cours des itérations sont distribuées selon $\phi^\star$. Une chaîne récurrente irréductible est dite *récurrente positive* si elle admet une distribution stationnaire. Si la probabilité de revenir infiniment souvent en tous points de l'espace vaut 1, alors la chaîne est dite *Harris récurrente*. Une chaîne *ergodique* est une chaîne irréductible, apériodique et Harris récurrente positive.

La loi forte des grands nombres s'applique aux chaînes ergodiques, c'est à dire que si $f = \int_{\mathcal{X}} g(x)dP(x)$ et que $(x_1, \ldots, x_m)$ est un échantillon de $P(x)$ obtenu grâce à une chaîne MCMC ergodique, alors $\bar{f_m} = \sum_{i=1}^{m} g(x_i)/m$ converge presque sûrement vers f pour m grand. Dans le cas de l'estimation du seuil optimal d'un marqueur, si la chaîne MCMC dont sont issus les échantillons de la distribution a posteriori des paramètres de distribution des marqueurs dans les deux groupes est ergodique, alors pour toute valeur seuil c, la fonction d'utilité moyenne sur l'ensemble des valeurs de paramètres échantillonnées converge presque sûrement vers la vraie valeur de la fonction d'utilité en ce point. Ceci ne veut pas dire pour autant que la valeur qui maximise la fonction d'utilité moyenne soit la valeur qui maximise la vraie fonction d'utilité. Pour montrer ce dernier point, il faut faire appel à la notion d'*hypoconvergence*, un type particulier de convergence utilisé dans les problèmes d'optimisation (Attouch, 1984 ; Geyer, 1994).

Supposons que $V(c, \boldsymbol{\theta})$ soit une fonction définie sur $\mathcal{C} \times \Theta$, où $\mathcal{C}$ est un espace métrique séparable complet, $(\mathcal{C}, \mathcal{B}, \phi)$ est un espace probabilisé complet, avec Θ un sous ensemble Borélien des espaces métriques séparables complets. On note :

$$\mathcal{U}_n^\star(c) = \frac{1}{m} \sum_{i=1}^{m} V(c, \boldsymbol{\theta}_i)$$

et on suppose que la chaîne MCMC utilisée pour échantillonner les valeurs $\boldsymbol{\theta}_i, i = 1, \ldots, m$ est ergodique. On suppose également que pour tout $\boldsymbol{\theta} \in \Theta$, la fonction V est semi-continue

inférieurement en tout point c pour toutes les valeurs de $\boldsymbol{\theta}$, à l'exception de quelques valeurs de $\boldsymbol{\theta}$ dépendant de ϕ et éventuellement de c, ainsi que la fonction est continue supérieurement en tout point c, à l'exception de quelques valeurs de $\boldsymbol{\theta}$ dépendant de ϕ mais ne pouvant pas dépendre de c cette fois-ci. Dans ce cas, $\mathcal{U}_m^\star$ épiconverge vers $\mathcal{U}^\star$ presque sûrement. La preuve de ce théorème est donnée dans Geyer (1994).

L'intérêt de l'hypoconvergence est que, si $\mathcal{C}$ est un espace compact, s'il existe un unique minimiseur de $\mathcal{U}^\star$, noté $c^\star$, et que $\hat{c}_m$ est une séquence d'ε-maximiseurs de V_m, c'est à dire tels que :

$$\mathcal{U}_m^\star(\hat{c}_m) \geq \sup_{c\in\mathcal{C}} \mathcal{U}_m^\star(c) - \varepsilon_m$$

avec $\varepsilon_m \to 0$, alors $\hat{c}_m$ tend en probabilité vers $c^\star$ et $\mathcal{U}_m^\star(\hat{c}_m)$ tend en distribution vers $\mathcal{U}^\star(c^\star)$. La preuve de ce théorème est donnée par Attouch (1984).

Ainsi, si les conditions présentées ci-dessus sont satisfaites, ce qui est vrai dans de nombreux cas, alors le seuil qui minimise la moyenne des fonctions d'utilité converge vers le seuil minimisant la vraie fonction d'utilité. Cette méthode est, entre autres, utilisée par Wang et Geisser (2005) pour estimer le seuil optimal d'un marqueur lorsque les a priori des paramètres de distribution des marqueurs dans les deux groupes ne sont pas conjugués. Néanmoins, aucun intervalle de confiance de l'estimation obtenue n'est fourni.

5.2.1.2 Intervalle de confiance du maximiseur de la moyenne des fonctions d'utilité

L'objectif de cette partie est de détailler les conditions nécessaires pour montrer que $\sqrt{m}(\hat{c}_m - c^\star)$ tend asymptotiquement vers une loi normale. La présentation est effectuée de manière générique, en considérant que $\hat{c}_m$ et $c^\star$ peuvent éventuellement être des vecteurs, lorsqu'il faut déterminer les valeurs optimales de plusieurs paramètres. Les hypothèses suivantes sont supposées valides :

1. le minimiseur de $\mathcal{U}^\star$ est unique et l'espace $\mathcal{C}$ contient un voisinage ouvert de $c^\star$ dans $\mathbb{R}$;
2. $\hat{c}_m$ converge en probabilité vers $c^\star$;
3. $\mathcal{U}^\star(c) = \mathrm{E}(V(c, \boldsymbol{\theta}))$ peut être différenciée deux fois sous le signe espérance ;
4. $B = \nabla^2 \mathcal{U}^\star(c^\star)$ est définie positive ;
5. $\sqrt{m}\nabla\mathcal{U}_m(c^\star)$ tend en distribution vers une loi normale, $\mathcal{N}(0, \sigma^2)$;
6. $\nabla^3 \mathcal{U}_m^\star(c)$ est uniformément bornée presque sûrement dans un voisinage de $c^\star$.

Dans ce cas, $\nabla^2 \mathcal{U}_m^\star(\hat{c}_m)$ converge en probabilité vers B et $\sqrt{m}(\hat{c}_m - c^\star) \xrightarrow{\mathcal{D}} \mathcal{N}(0, B^{-1}VB^{-1})$. Ici, ∇ dénote le gradient, le vecteur des dérivées premières partielles ; ∇^2 correspond au Hessien, la matrice des dérivées secondes partielles.

La plupart de ces conditions sont similaires à celles utilisées pour démontrer la convergence asymptotique des estimateurs du maximum de vraisemblance vers une loi normale. La dernière condition peut être montrée en trouvant une fonction dominante et en utilisant le théorème de convergence dominée qui s'applique également aux chaînes MCMC. B peut être estimée par $\nabla^2 \mathcal{U}_m^\star(\hat{c}_m)$. La seule condition inhabituelle est la cinquième condition, qui est une extension du théorème central limite au cas des chaînes MCMC. Normalement, le théorème central limite ne s'applique que lorsque les échantillons sont indépendants, ce qui n'est pas le cas des échantillons issus d'une chaîne MCMC. Néanmoins, si la chaîne MCMC est géométriquement ou uniformément ergodique, alors le théorème central limite s'applique également aux chaînes MCMC (Gilks et *al.*, 1996). Ces deux dernières propriétés caractérisent la vitesse de convergence de la chaîne MCMC ; elles sont très difficiles à démontrer, même pour les chaînes MCMC les plus simples. De même, on notera qu'il est relativement difficile d'obtenir une estimation de σ, puisque les échantillons ne sont pas indépendants. Deux des principales méthodes en vue d'y parvenir sont les méthodes des moyennes par lot et des estimateurs par fenêtre (Gilks et *al.*, 1996).

5.2.1.3 Les limites de la méthode de maximisation de la moyenne des fonctions d'utilité

L'objectif initial était d'optimiser une fonction faisant intervenir plusieurs intégrales sur des variables aléatoires, ces variables correspondant aux paramètres de la distribution des marqueurs dans les deux groupes. En échantillonnant des valeurs dans la distribution a posteriori de ces paramètres grâce à une chaîne MCMC, il est possible de calculer l'utilité espérée par la méthode d'intégration de Monte Carlo ; sous des conditions rencontrées fréquemment, le maximiseur de la fonction ainsi approximée converge vers le maximiseur de la vraie fonction. Ceci correspond aux méthodes de maximisation de Monte Carlo. Elles sont couramment utilisées dans les problèmes d'optimisation complexes (Kall, 1986).

Néanmoins, il est très rare qu'un intervalle de confiance soit fourni pour ces types de problèmes, alors que cela est indispensable pour l'estimation du seuil optimal d'un marqueur. Les conditions nécessaires pour que l'estimateur du seuil suive une loi normale sont très difficiles à montrer, même dans les cas les plus simples. De plus, la validité de l'intervalle de confiance ainsi

construit n'est qu'asymptotique. Ainsi, cette méthode n'est pas très adaptée pour la construction de l'intervalle de confiance de l'estimation du seuil optimal d'un marqueur. Une autre piste à été explorée, consistant à retenir la moyenne des seuils optimaux de chacune des fonctions d'utilité. Cette seconde méthode est présentée ci-après.

Avant de terminer cette partie, on notera que le terme optimisation de Monte Carlo est aussi employé pour décrire des méthodes d'exploration aléatoire de la fonction d'utilité, notamment par recuit simulé (Muller, 1999), mais ces méthodes ne quantifient pas le degré d'imprécision des estimations ; elles ne sont donc pas utilisables lorsqu'un intervalle de confiance est désiré.

5.2.2 Moyenne des maximums des fonctions d'utilité

Dans le cas de l'estimation du seuil optimal d'un marqueur, la fonction d'utilité est donnée par :

$$\mathcal{U}^{\star}(c) = \int_{\Theta_0}\int_{\Theta_1} (P(\tilde{y}_1 > c|\boldsymbol{\theta}_1) + P(\tilde{y}_0 \leq c|\boldsymbol{\theta}_0) \times R) P(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1|\mathbf{y})\, d\boldsymbol{\theta}_0 d\boldsymbol{\theta}_1 \tag{5.5}$$

$P(\tilde{y}_1 > c|\boldsymbol{\theta}_1)$ est la probabilité prédite qu'une valeur de marqueur chez un malade soit supérieure au seuil c ; $P(\tilde{y}_0 \leq c|\boldsymbol{\theta}_0)$ correspond à celle qu'une valeur de marqueur chez un non malade soit inférieure ou égale au seuil c.

Supposons qu'une chaîne MCMC permette d'échantillonner m valeurs dans la distribution a posteriori des paramètres de distribution du marqueur chez les malades et les non malades d'après les mesures $\mathbf{y}$ qui ont été effectuées dans les deux groupes. Alors la fonction (5.5) peut être approximée par :

$$\mathcal{U}^{\star}_m(c) = \frac{1}{m}\sum_{i=1}^{m} P(\tilde{y}_1 > c|\boldsymbol{\theta}_{1i}) + P(\tilde{y}_0 \leq c|\boldsymbol{\theta}_{0i}) R \tag{5.6}$$

Chaque itération de la chaîne MCMC conduit à une fonction d'utilité. Les fonctions d'utilité obtenues sur les m itérations constituent la distribution a posteriori de la fonction d'utilité. A chaque itération, il est possible de calculer la valeur de seuil qui maximise la fonction d'utilité de l'itération en question, soit par une formule explicite, soit par une méthode de type Newton-Raphson. Les seuils obtenus sur les m itérations constituent la distribution a posteriori du seuil optimal. Le mode, la moyenne ou la médiane des valeurs correspondent à des estimations ponctuelles du seuil optimal ; un intervalle de crédibilité est constructible par la méthode des quantiles ou par la méthode HDP (annexe A).

La distribution a posteriori du seuil optimal est obtenue alors qu'aucune distribution a priori n'a été fixée pour ce dernier. En réalité, une fois le ratio bénéfice net sur coût net et la prévalence fixés, le seuil optimal est entièrement déterminé par la donnée de la sensibilité et de la spécificité du marqueur ; or la sensibilité est directement obtenue à partir des paramètres de distribution du marqueur chez les malades ; de même, la spécificité est obtenue à partir des paramètres de distribution chez les non malades. Ici, les seules variables aléatoires sont donc les paramètres de distribution du marqueur dans les deux groupes. L'information a priori utilisée pour caractériser la distribution de ces variables se transmet par des liens logiques au seuil optimal. La distribution a priori du seuil est donc fixée par l'information a priori attribuée aux paramètres de distribution du marqueur.

Cette méthode de détermination du seuil optimal et de son intervalle de crédibilité est très simple à utiliser une fois que des valeurs ont été échantillonnées dans la distribution a posteriori des paramètres de distribution du marqueur. Les seules contraintes sont donc de savoir échantillonner dans la distribution a posteriori de ces paramètres, ainsi que de connaître la distribution du marqueur dans les deux groupes. La méthode proposée sera appelée par la suite méthode Bayésienne d'estimation du seuil et de son intervalle de confiance ; celle qui repose sur la maximisation de la moyenne des fonctions d'utilité sera appelée méthode de Wang et Geisser (WG). La figure 5.4 représente schématiquement la différence entre ces deux méthodes :

- l'une calcule les fonctions d'utilité, en détermine la moyenne, puis estime la valeur qui maximise cette fonction moyenne (méthode de Wang et Geisser) ;
- l'autre calcule les fonctions d'utilité, détermine les valeurs de seuil optimal pour chacune d'entre elles, puis retient la moyenne – ou le mode ou encore la médiane – des seuils obtenus (méthode Bayésienne d'estimation du seuil).

5.3 Estimation du seuil optimal pour un marqueur dynamique

5.3.1 Méthode et application aux données de PSA

Dans le cas d'un marqueur dynamique, les valeurs mesurées $\mathbf{y}$ ne correspondent pas directement au marqueur. Ces valeurs de marqueur, par exemple le nadir de PSA, sont échantillonnées dans la distribution a posteriori du nadir de chaque patient d'après les valeurs de PSA. A chaque itération de la chaîne MCMC, les valeurs de nadir de PSA échantillonnées chez les malades suivaient une loi log normale de paramètre $\boldsymbol{\theta}_{1i} = \{\mu_{1i}, \sigma^2_{1i}\}$, ainsi que chez les non malades, de

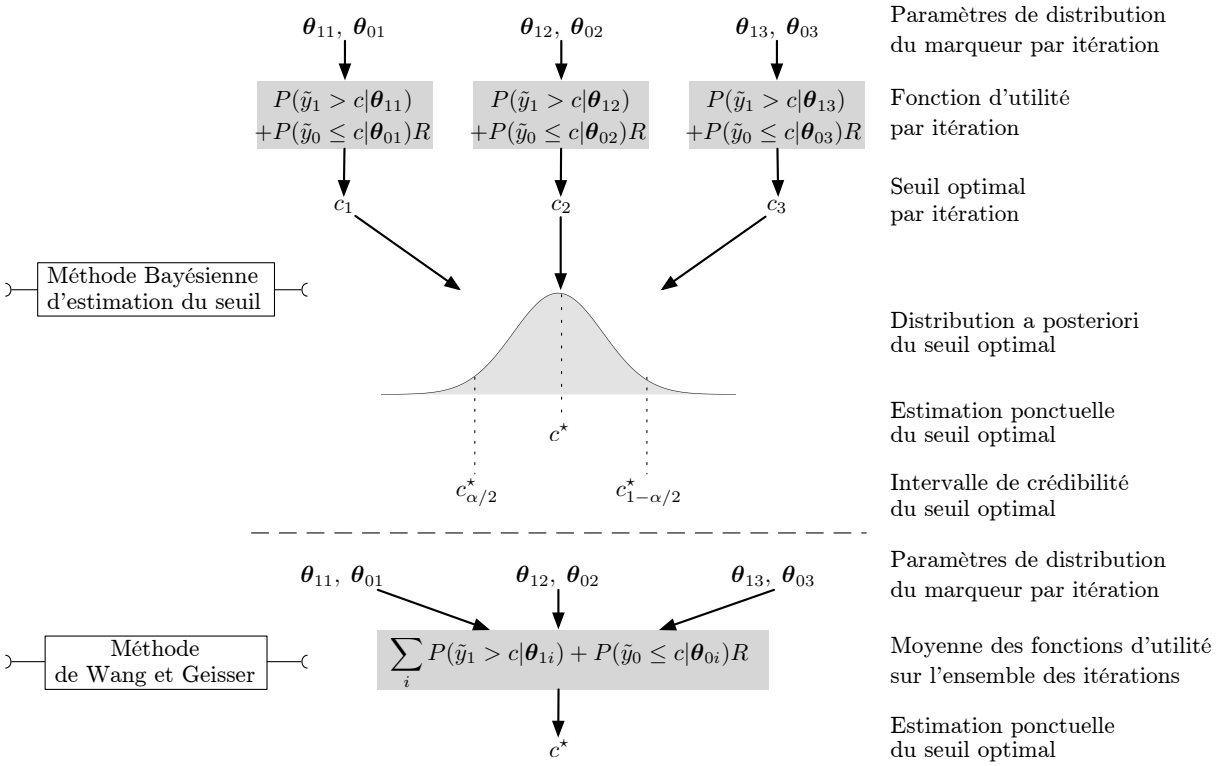

Figure 5.4 – Comparaison de la méthode Bayésienne d'estimation du seuil et de la méthode de Wang et Geisser.

paramètre cette fois $\boldsymbol{\theta}_{0i} = \{\mu_{0i}, \sigma^2_{0i}\}$. En utilisant des a priori non informatifs pour ces paramètres ($P(\mu_j, \sigma^2_j) = 1/\sigma^2_j, j = 0, 1$), une valeur de chaque paramètre a été échantillonnée à chaque itération dans la distribution a posteriori ; celle-ci est donnée pour une itération i par :

$$1/\sigma^2_{ji}|\text{nadir}_{ji} \hookrightarrow \frac{\chi^2_{n_j-1}}{\sum_{k=1}^{n_j}\left(\ln(\text{nadir}_{jik}) - \overline{\ln(\mathbf{nadir}_{ji})}\right)^2} \quad \mu_{ji}|\mathbf{nadir}_{ji} \hookrightarrow \mathcal{N}\left(\overline{\ln(\mathbf{nadir}_{ji})}, \sigma^2_{ji}/n_j\right)$$

avec $\overline{\ln(\mathbf{nadir}_{ji})} = \sum_{k=1}^{n_j} \ln(\text{nadir}_{jik})$. $\mathbf{nadir}_{ji}$ correspond aux n_j valeurs de nadirs du groupe j échantillonnées à l'itération i de la chaîne MCMC. La sensibilité et la spécificité du nadir de PSA ont ainsi pu être calculées, permettant à chaque itération la construction de la fonction d'utilité et la détermination du seuil optimal de l'itération en question. Les 3000 valeurs de seuils échantillonnées constituaient la distribution a posteriori du seuil optimal du logarithme du nadir de PSA.

Pour le groupe des malades, pour des valeurs de logarithme de nadir supérieures à 1, l'ajustement entre une loi log normale et les valeurs de nadir échantillonnées était moins adéquat. En réalité, l'intersection entre les courbes de densité de probabilité chez les non malades et les

malades avait lieu en dessous de 1 sur l'échelle du logarithme du nadir. Ainsi, il fallait accorder beaucoup d'importance à la modélisation correcte de la distribution des valeurs de nadirs en dessous de 1 chez les malades, la zone au dessus étant moins importante. L'adéquation entre la distribution des logarithmes de nadirs chez les malades à chaque itération et une loi normale a donc été testée à partir d'un test d'adéquation partielle, pour des valeurs inférieures à 1. Hand (1997) recommande d'ailleurs d'accorder beaucoup d'importance à la modélisation des distributions dans les zones où elles se recoupent dans les deux groupes, car ce sont dans ces régions que les problèmes se posent.

La méthode présentée ci-dessus tient compte, dans l'estimation des seuils, de l'incertitude quant à l'estimation des valeurs de nadirs de PSA, puisque ces valeurs sont échantillonnées dans leur distribution a posteriori d'après les mesures de PSA. Ainsi, l'estimation ponctuelle et l'intervalle de crédibilité du seuil optimal tiennent compte du fait que les marqueurs ne sont pas mesurés, mais estimés avec une certaine précision. Ceci n'aurait pas été directement possible avec les méthodes d'estimation de seuil existant actuellement.

5.3.2 Un modèle, deux paramétrisations

Le modèle utilisé pour construire les profils de PSA des patients (équation 4.7) est paramétré en termes de valeurs d'effets aléatoires $\{r_1, r_2, r_3, r_4\}$, en utilisant des distributions souples pour les effets aléatoires r_1 et r_4 ; ce modèle est appelé par la suite modèle 1. Il est représenté de manière simplifiée sur la figure 5.5. Pour ne pas alourdir la figure, la variance des résidus est représentée comme étant constante, même si ce n'est pas le cas pour le vrai modèle implémenté. C'est à partir des estimations de $\{r_1, r_2, r_3, r_4\}$ par patient que le nadir a été ensuite calculé.

En réalité, étant donné que l'analyse a porté principalement sur le nadir de PSA, il aurait été utile de construire un modèle faisant intervenir directement dans ses paramètres le nadir des patients, en utilisant des distributions souples pour le nadir et les autres paramètres (figure 5.6). Ce modèle est appelé par la suite modèle 2. C'est un modèle purement conceptuel, car il n'est pas possible d'obtenir un modèle équivalent au modèle 1, en termes de courbes prédites, en introduisant le nadir comme un des paramètres. Des distributions souples étant utilisées dans les deux modèles pour décrire les distributions des paramètres sur l'ensemble des patients, les distributions a posteriori des nadirs calculés à partir des paramètres $\{r_1, r_2, r_3, r_4\}$ du modèle 1

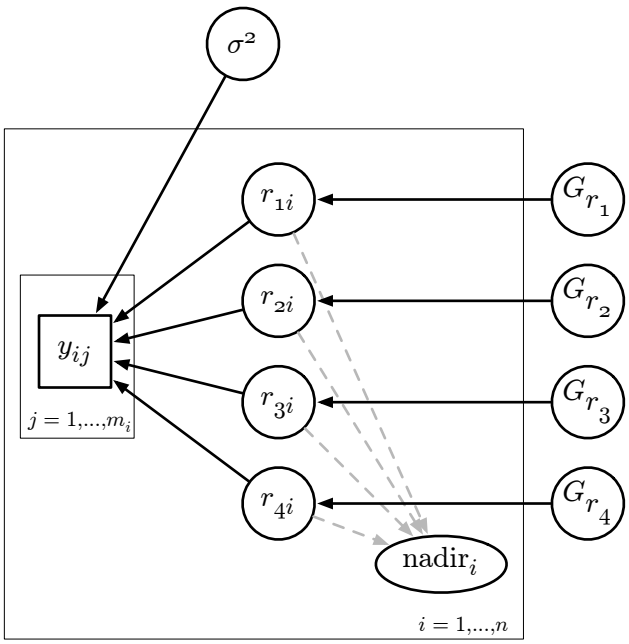

Figure 5.5 – Représentation schématique du modèle 1 (G : distributions de paramètres sur l'ensemble des patients).

ou directement échantillonnés pour le modèle 2 devraient être similaires. Ainsi, par la suite, les modèles 1 et 2 sont considérés comme équivalents d'un point de vue conceptuel.

L'estimation du seuil optimal nécessite de pouvoir échantillonner dans la distribution a posteriori des paramètres de distribution du nadir dans les deux groupes. Ainsi, il serait utile d'échantillonner les valeurs de nadir dans deux distributions selon le statut des patients et donc d'introduire ce statut ainsi que des paramètres de distribution du marqueur pour les deux groupes (μ_0, σ_0^2 et μ_1, σ_1^2). Ceci correspond au modèle 3 (figure 5.7). Encore une fois, d'un point de vue technique, ce modèle n'est pas constructible, il n'est donc considéré que d'un point de vue conceptuel. Ce dernier modèle permettrait d'obtenir directement des échantillons de valeurs issues de la distribution a posteriori des paramètres de distribution du marqueur dans les deux groupes et non de les recalculer après coup comme avec le modèle 1. Si la paramétrisation du modèle 2 (et donc du modèle 1) est suffisamment souple au niveau de la distribution du nadir, alors la distribution ainsi obtenue doit reconstituer la distribution des nadirs du modèle 3, constituée d'un mélange de deux distributions selon le statut des patients (M), et ceci, bien que le statut des patients n'ait pas été introduit dans le modèle 2. Ainsi, dans la chaîne MCMC, l'échantillonnage des valeurs de nadir par patient selon G_{nadir} et les données de PSA (modèle 2), ou selon le statut des patients, μ_1, μ_0, σ_1, σ_0 et les données de PSA (modèle 3) doit être

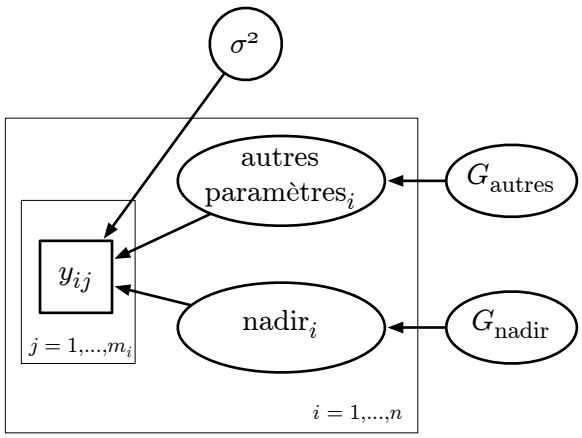

Figure 5.6 – Représentation schématique du modèle 2 (G : distributions de paramètres sur l'ensemble des patients).

équivalent. Sous l'hypothèse que les valeurs de nadir dans les deux groupes suivent bien des distributions log normales, les distributions des valeurs de paramètres $\mu_1,\ \mu_0,\ \sigma_1$ et σ_0 calculées à partir du modèle 1, ou échantillonnées directement dans la chaîne MCMC du modèle 3, doivent être équivalentes.

La seule limite à la discussion précédente est que le modèle 3 contraint les valeurs de nadir à suivre des lois log normales par groupe, alors que le modèle 2, et donc le modèle 1, n'introduisent aucune contrainte. Néanmoins, les valeurs de nadir issues du modèle 1 semblaient suivre des distributions log normales à chaque itération ; ainsi, même sans contrainte, les résultats du modèle 1 devraient êtres équivalents à ceux du modèle 3.

Dans une démarche où l'objectif est de déterminer le marqueur – reflétant la cinétique d'un biomarqueur – le plus approprié pour effectuer le diagnostic d'une maladie ou le pronostic d'une typologie d'évolution, il semble plus logique de ne pas introduire dans un premier temps le statut du patient dans la modélisation des profils, ni de contraindre la distribution des paramètres selon ce statut.

5.4 Article accepté dans le Biometrical Journal

L'article accepté dans le Biometrical Journal présente un ensemble de simulations réalisées dans le cas d'un marqueur reflétant la cinétique d'un biomarqueur, afin de mesurer le biais lié à cette méthode d'estimation du seuil, ainsi que la probabilité de couverture de l'intervalle de crédibilité construit.

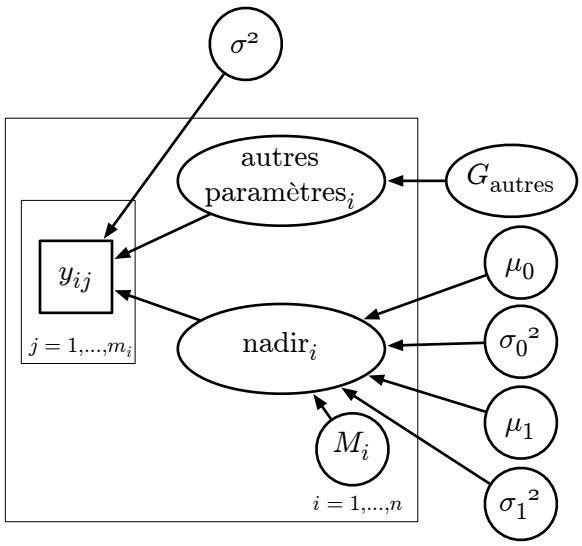

Figure 5.7 – Représentation schématique du modèle 3 (G : distributions de paramètres sur l'ensemble des patients).

Les simulations ont été conduites en faisant varier un certain nombre de paramètres, dont l'écart-type de la distribution des marqueurs chez les malades et les non malades et le ratio bénéfice net sur coût net. Comme indiqué dans la partie 5.1.1, le problème de l'estimation du seuil optimal d'un marqueur correspond au problème de l'estimation du point d'intersection entre deux courbes. Ces courbes correspondent aux densités de probabilité de valeurs de marqueur chez les malades et les non malades, la courbe des malades étant multipliée par α et celle des non malades par β, α et β étant choisis tels que :

$$\frac{\alpha}{\beta} = \frac{\pi}{1-\pi} \times \frac{BN}{CN}$$

Ainsi, le fait de modifier la prévalence ou le ratio bénéfice net sur coût net revient à modifier la forme de ces courbes. Le seuil optimal pour une prévalence $\pi = 1/2$ et $BN/CN = 2$ est identique à celui obtenu pour $\pi = 3/4$ et $BN/CN = 2/3$. L'objectif des simulations est de tester la capacité des méthodes proposées à estimer correctement le point d'intersection des courbes et son intervalle de crédibilité pour une grande variété de formes de courbes.

Trois paramètres influencent la forme de ces courbes : l'écart-type de la distribution du marqueur chez les malades et les non malades, la prévalence de la maladie et le ratio bénéfice net sur coût net. Afin de limiter le nombre de simulations réalisées, la prévalence a été fixée à 0,5. De même, l'étendue des valeurs retenues pour les trois autres paramètres était faible, l'essentiel

étant que la combinaison des effets des différents paramètres sur les courbes entraîne une grande variété de formes de courbes pour l'estimation du seuil optimal.

L'article présente également l'application de cette méthode aux données de PSA après traitement UFHI, avec l'estimation du seuil optimal de nadir de PSA pour discriminer les patients selon le résultat attendu des biopsies.

Les distributions conditionnelles des paramètres du modèle utilisé lors des simulations sont données dans l'annexe C.

5.4.1 Article

bimj header will be provided by the publisher

seddate30 November 2004

A Bayesian Method to Estimate the Optimal Threshold of a Longitudinal Biomarker

Fabien Subtil[*1,2] and **Muriel Rabilloud**[1,2]

[1] Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

[2] Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France

Received 15 November 2004, accepted 2 December 2004

Published online 3 December 2004

Summary

The objective of this study was to develop methods to estimate the optimal threshold of a longitudinal biomarker and its credible interval when the diagnostic test is based on a criterion that reflects a dynamic progression of that biomarker. Two methods are proposed: one parametric and one non-parametric. In both cases, the Bayesian inference was used to derive the posterior distribution of the optimal threshold from which an estimate and a credible interval could be obtained. A numerical study shows that the bias of the parametric method is low and the coverage probability of the credible interval close to the nominal value, with a small coverage asymmetry in some cases. This is also true for the non-parametric method in case of large sample sizes. Both methods were applied to estimate the optimal Prostate-specific antigen nadir value to diagnose prostate cancer recurrence after a High-Intensity Focused Ultrasound treatment. The parametric method can also be applied to non-longitudinal biomarkers.

* Corresponding author: e-mail: fabien.subtil@chu-lyon.fr, Phone: +33 (0)4 7211 5751 Fax: +33 (0)4 7211 5141

Key words: Diagnostic markers, Longitudinal study, Optimal cut-point, Prostate Specific Antigen, Sensitivity and Specificity.

1 Introduction

Medical decision making has been revolutionized by the use of biomarkers as screening or diagnostic tests, or for patient monitoring. Biomarkers have been developed in most medical specialties, especially oncology (Schiffer, 2009). Their use as diagnostic tests requires the assessment of their diagnostic accuracy. Evaluation and comparison of diagnostic accuracies of several quantitative biomarkers are currently made using Receiver Operating Characteristic (ROC) analyses. The area under the ROC curve –the plot of sensitivity against one minus specificity across all possible threshold values for the biomarker– is indeed a measure of the global accuracy of a marker (Pepe, 2003), regardless of the threshold above or below which the test is judged positive. After selection of the best biomarker in terms of diagnostic accuracy through area under ROC curves, a threshold, later called the optimal threshold, is to be chosen for disease diagnosis. This threshold is the one that maximises a utility function; for example, the threshold that maximises the number of well-classified patients. In clinical articles, little information is given on the way the optimal threshold is estimated, and its credible interval is rarely mentioned.

There are three broad categories of methods to estimate the optimal threshold of a biomarker. Parametric ones rely on the assumption that the biomarker follows a given distribution in diseased and non-diseased subjects; from these distributions, the optimal threshold can be estimated using, in some cases, an explicit formula, or numerical optimisation. Methods have been proposed for gamma, normal, lognormal distributions, or normal distributions obtained after a Box-Cox transformation (Fluss et al., 2005; Schisterman and Perkins, 2007). When an explicit formula of the optimal threshold exists, the confidence interval can be calculated by the delta method; otherwise, by bootstrap (Schisterman and Perkins, 2007). Ruopp et al. (2008) have proposed a semi-parametric ROC-GLM method essentially based on a parametric smoothing of the empirical ROC curve from which the threshold can be estimated. The most common non-parametric approach uses the empirical cumulative distribution functions as estimates of the cumulative distribution

function of the marker in diseased and non-diseased subjects, but this approach has higher bias and root mean squared error with respect to other methods (Fluss et al., 2005). Some methods rely on a kernel smoothing of the cumulative distribution function (Fluss et al., 2005) or on the smoothing of the ROC curve with a loess procedure (Leeflang et al., 2008). In most cases, there is no explicit formula for the optimal threshold; thus, its confidence interval must be estimated using bootstrap. Several bootstrap methods have been proposed but, in some cases, the coverage probability may be far from the nominal one (Schisterman and Perkins, 2007). There is thus a need for new methods to estimate the optimal threshold of a biomarker used to diagnose an already existing disease.

In some diseases, longitudinal measurements of a biomarker have become commonplace. Examples are monitoring of CD4+ cells in patients with HIV (Zheng and Heagerty, 2007), monitoring of the PSA to diagnose recurrence of prostate cancer (Subtil and Rabilloud, 2010), or the follow-up of renal transplanted patients with PCR measurements to predict cytomegalovirus disease (Subtil et al., 2009). In longitudinal measurements, the biomarker value by itself is not necessarily the best criterion for the diagnosis; criteria reflecting the dynamic progression of the biomarker –such as a slope– may be more accurate. In prostate cancer, the smallest PSA value measured after a treatment, called the nadir, has been found to be the best dynamic criterion to diagnose recurrence (Subtil and Rabilloud, 2010). In the present study, PSA measurements were modelled using a non-linear mixed model and, in each patient, the nadir was calculated from the parameters of this model. The assessment of the diagnostic accuracy was based on the nadir values issued from the model. However, no threshold has been defined allowing the use of the nadir as a diagnostic test. The aforementioned methods that estimate the optimal threshold and its confidence interval are not directly applicable when the criterion is not directly measured but only estimated from a model.

These remarks have motivated the development of a new method to estimate the optimal threshold of a biomarker and its credible interval. This method does not call for an explicit formula of the optimal threshold to derive its credible interval and can be adapted to the case of longitudinal biomarkers. Numerical studies were conducted to analyse the bias of the estimated threshold and the coverage probability of

its credible interval in the case of a longitudinal biomarker. The method was then applied to estimate the threshold of the PSA nadir when this nadir is used for the diagnosis of prostate cancer recurrence after a High-Intensity Focused Ultrasound (HIFU) treatment.

2 Methods

The method proposed to estimate the optimal threshold of a biomarker and its credible interval relies on Bayesian inference. A thorough presentation of Bayesian inference can be found in Gelman et al. (2004), we just evoke here some features essential to introduce the method. Suppose we have some prior knowledge on the values of the parameters of a model. This knowledge may be characterized by prior distributions. New measurements are used to update these prior distributions and derive the corresponding posterior distribution of the parameters. If a specific parameter is a combination or a function of some other parameters, then its posterior distribution can be derived from the posterior distribution of the other ones (Gelman et al., 2004). Here, the posterior distribution of the optimal threshold will be derived from the posterior distribution of the parameters of the biomarker distributions in diseased and non-diseased subjects. Before describing the method, we specify what is called "the optimal threshold".

2.1 Decision-theoric optimal threshold

One approach to determine the optimal threshold of a quantitative biomarker is to maximise its expected utility within the population to test, which is a function of the sensitivity and specificity of the test, the prevalence of the disease, the net benefit (NB) of treating a diseased subject, and the net cost (NC) of treating a healthy subject (Jund et al., 2005; Sox et al., 1988). These benefits and costs are quantified in terms of utility, a measure of the state of health or preference for a state of health in various settings (such as life expectancy or quality of life). Depending on the result of the test and the patient true status, four specific utilities can be defined, U_{FN}, U_{TP}, U_{FP}, and U_{TN}, to assess the state of health of a

diseased-untreated subject (false negative), a diseased-treated one (true positive), a healthy subject mistakenly treated (false positive), and a healthy untreated one (true negative), respectively. Denoting $\mathrm{Sen(c)}$ and $\mathrm{Spe(c)}$ the sensitivity and specificity of the test associated to a threshold c, π the prevalence of the disease, and N the size of the whole population, the expected utility of the test over the population for the threshold c, denoted $U(c)$, is defined by:

$$U\left(c\right) = N\left[\mathrm{Sen}\left(c\right)\pi U_{TP} + \left\{1 - \mathrm{Spe}\left(c\right)\right\}\left(1-\pi\right)U_{FP} + \mathrm{Spe}\left(c\right)\left(1-\pi\right)U_{TN} + \left\{1 - \mathrm{Sen}\left(c\right)\right\}\pi U_{FN}\right]$$

It can be rewritten:

$$U\left(c\right) = D\left(\mathrm{Sen}\left(c\right) + \mathrm{Spe}\left(c\right)R + G\right)$$

with $D = NB \times \pi \times N$, $R = (NC/NB) \times (1-\pi)/\pi$, $G = ((1-\pi)U_{FP} + \pi U_{FN})/(NB \times \pi)$, $NB = U_{TP} - U_{FN}$ and $NC = U_{TN} - U_{FP}$.

Discarding all terms that do not depend on the threshold, the optimal threshold is the one that maximises the function:

$$\tilde{U}\left(c\right) = \mathrm{Sen}\left(c\right) + \mathrm{Spe}\left(c\right)R \tag{1}$$

The ratio NB/NC can be interpreted as the number of healthy patients a clinician would accept to treat mistakenly in order not to leave a diseased subject without treatment (DeNeef and Kent, 1993).

Several studies have defined the optimal threshold as the one that maximises the function $\mathrm{Sen(c)} + \mathrm{Spe(c)}$, called the Youden index (Fluss et al., 2005; Schisterman and Perkins, 2007). This is a subcase of the above-cited utility function, considering the value of R equal to one; this is the case, for example, when the prevalence is 0.5 and the net benefit equal to the net cost.

When the prevalence and the distributions of the biomarker are known in diseased and non-diseased subjects, the sensitivity and the specificity, and then the utility function, can be easily estimated for all thresholds.

2.2 Bayesian estimate of the optimal threshold and its credible interval for biomarker

We consider first the case of biomarkers that are used to make the diagnosis of an already existing disease. Let $\boldsymbol{\theta}_i$ denote the parameter(s) of the distribution function of the biomarker, $i = 0$ for non-diseased subjects and $i = 1$ for diseased ones ($\boldsymbol{\theta}_i$ might be a vector of parameters), and let $p(\boldsymbol{\theta}_i)$ denote the prior distribution of $\boldsymbol{\theta}_i$. The prior distribution may be non informative if no prior information is available. Let $\mathbf{x}_i = \{x_{i1}, \ldots, x_{in_i}\}$ be the set of n_i biomarker measurements in population i (diseased or non-diseased). The posterior distribution of $\boldsymbol{\theta}_i$ is obtained from Bayes theorem:

$$p(\boldsymbol{\theta}_i|\mathbf{x}_i) \propto p(\mathbf{x}_i|\boldsymbol{\theta}_i)\, p(\boldsymbol{\theta}_i)$$

This posterior distribution is not always known explicitly; then, algorithms are used to obtain a sample of the parameters from their posterior distributions, such as MCMC algorithms (Gelman et al., 2004).

Suppose, for example, that the biomarker values follow a Gaussian distribution with mean μ_i and variance σ_i^2; in this case, $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2)$. Under the non-informative prior $p(\mu_i, \sigma_i^2) \propto \sigma_i^{-2}$, the marginal posterior distribution of σ_i^2 is $(\mathbf{x}_i - \bar{\mathbf{x}}_i)^T (\mathbf{x}_i - \bar{\mathbf{x}}_i) \Big/ \chi^2_{n_i-1}$, and the conditional distribution of μ_i given σ_i^2 is normal: $N(\bar{\mathbf{x}}_i, \sigma_i^2/n_i)$. In this case, it is easy to sample K values from the posterior distribution of these parameters and then obtain K sample functions from the posterior distribution of the utility function:

$$\tilde{U}_k(c) = 1 - \Phi((c - \mu_{1k})/\sigma_{1k}) + R \times \Phi((c - \mu_{0k})/\sigma_{0k})$$

where Φ denotes the standard normal distribution function. The value of c that maximises each utility function (denoted $c_k^\star,\ k = 1, \ldots, K$) can be calculated using an optimisation algorithm such as the

Newton-Raphson one. In the case of normal distributions in both populations, an explicit solution exists (Schisterman and Perkins, 2007):

$$c_k^\star = \frac{\mu_{0k}\left(b_k^2-1\right) - a_k + b_k\sqrt{a_k^2 + \left(b_k^2-1\right)\sigma_{0k}^2 \log\left(b_k^2 R^2\right)}}{b_k^2-1}, \text{ with } a_k = \mu_{1k} - \mu_{0k} \text{ and } b_k = \frac{\sigma_{1k}}{\sigma_{0k}}$$

or, when the standard deviations in the two groups are equal:

$$c_k^\star = \frac{2\sigma_k^2 \log\left(R\right) + \mu_{1k}^2 - \mu_{0k}^2}{2a_k}$$

where σ_k denotes the standard deviation common to the two groups.

The previous K values of $c^\star$ are a sample of the posterior distribution of the optimal threshold. An estimate can be obtained using the mode, the median, or the mean of these values, and a $1-\alpha$ credible interval using either the highest posterior density (HPD) region (Tanner, 1996), called below the HPD method, or the empirical $\alpha/2$ and $1-\alpha/2$ quantiles of the K values, called below the quantile method.

In their article, Wang and Geisser (2005) proposed not to calculate a threshold for each utility function but to calculate the one that maximises the mean of all utility functions. Using arguments of hypoconvergence, it was demonstrated that this estimate tends, in probability, to the true optimal threshold. A credible interval could be obtained using the central limit theorem applied to Markov chains (Chen et al., 1999), but the type of convergence of the MCMC chain required to validate this method are either not fulfilled by most MCMC chains or too difficult to prove. This is why the method based on the posterior distribution of the optimal threshold has been favoured.

2.3 Bayesian parametric estimation of the optimal threshold and its credible interval for dynamic diagnostic criteria

In the case of longitudinal measurements, a criterion that reflects the dynamic progression of a marker, called below the marker trajectory, may be used as diagnostic test instead of the marker values themselves.

However, this criterion is calculated from the marker values. When the criterion is sensitive to the measurement time or to outlying biomarker measurements, it may be better to model the biomarker values over time and calculate the criterion from the modelled trajectories than to use the criterion values calculated directly from the marker measurements. Simulation results supporting this idea are given as supplementary material. In this case, careful attention should be given to the modelling of the trajectories, especially relaxing the assumptions that are too rigid for the data. In the present article, a mixed model was used to model the biomarker trajectory because of the flexibility of such an approach in modelling the covariance structure between and within observations (Laird and Ware, 1982).

Suppose the l^{th} of the N patients has m_l longitudinal measurements denoted y_{lj}, $j = 1, \ldots, m_l$, taken at time t_{lj}. The longitudinal model is given by:

$$y_{lj} = \psi\left(t_{lj}, \boldsymbol{\eta}_l\right) + \varepsilon_{lj} \quad \varepsilon_{lj} \sim N\left(0, \sigma_\varepsilon^2\right)$$

where ψ is the biomarker trajectory function that depends on the measurement time, and on the random effects $\boldsymbol{\eta}_l$. The trajectory function may not be linear. It is assumed that the dynamic criterion can be calculated for each patient by a function of its random effects; e.g., $x_l = g\left(\boldsymbol{\eta}_l\right)$. By combining the prior distributions of the model parameters and the likelihood of the measurements, it is possible to derive the posterior distributions of these parameters, especially the posterior distributions of random effects, and, consequently, that of the dynamic criterion for each patient. From these posteriors, K criterion values can be sampled for each patient, leading to K distributions of the criterion in the diseased and non-diseased subjects.

One difficulty is to identify a common distribution suitable for each iteration in diseased and non-diseased subjects, with parameters $\boldsymbol{\theta}_i$. Assuming then prior distributions for the $\boldsymbol{\theta}_i$ –possibly non-informative priors–, the corresponding posterior distributions can be derived, with K sampled values. The method to estimate the optimal threshold and its credible interval is now similar to the one used in section 2.2.

Other methods than mixed models may be used for longitudinal measurements; the only requirement to estimate the optimal threshold is to be able to sample from the posterior distribution of the diagnostic criterion for each patient; this is straightforward using a mixed model in a Bayesian context.

A detailed example on the way to sample from the posterior distribution of the optimal threshold is given in the Appendix.

2.4 Bayesian non-parametric estimation of the optimal threshold and its credible interval for dynamic diagnostic criteria

The major constraint in the previous section was that the diagnostic criterion is assumed to arise from a parametric distribution in the diseased and non-diseased groups, but this constraint can be relaxed. Let $\mathbf{x}_k$ denotes the vector of diagnostic criterion values obtained at iteration k for the N patients. One can calculate the utility function (1) from the empirical estimates of sensitivity and specificity obtained with the $\mathbf{x}_k$ values at iteration k. The optimal threshold at iteration k is the one, among the $\mathbf{x}_k$ values, that maximises the utility function.

The values of optimal threshold obtained for the K iterations forms the posterior distribution of the optimal threshold, from which a point estimate and credible interval can be obtained as in section 2.2.

3 Simulation Study

A simulation study was conducted to assess the properties of the proposed estimation method on a diagnostic criterion derived from a longitudinal biomarker. It is supposed that the measurements begin after a specific treatment, as it will be the case in the application section.

3.1 Design of the simulations

Let us consider N patients, of whom $N/2$ are developing a disease and $N/2$ are not. All are going to have a marker measurement just after the treatment, then every five days during the first twenty days,

then every twenty days until the hundredth day. Because those measurements are often delayed, the exact sampling times were sampled from a uniform distribution whose left boundary corresponded to the day the measurement should have been done and the right boundary to the day the next measurement should be done. The last day of follow-up was drawn, for each patient, from a Weibull distribution, with shape 9 and scale 100. Biomarker values were generated using the following model:

$$\log(y_{lj}) = \log\left(\exp(\eta_{1l}) + \exp(\eta_{2l} t_{lj})\right) + \varepsilon_{lj}, \quad \varepsilon_{lj} \sim N\left(0, 0.005^2\right)$$

$$\eta_{1l} \sim N\left(1.2, 0.003^2\right) \quad \eta_{2l} \sim \begin{cases} N\left(-0.3, \sigma_0^2\right) & \text{for controls;} \\ N\left(-0.1, \sigma_1^2\right) & \text{for cases.} \end{cases}$$

A small decrease in the biomarker value –i.e., a small value of η_2– was hence indicative of disease. Here, the diagnostic criterion was directly one of the random parameters of the model (η_2).

A MCMC algorithm was used to sample from the posterior distribution of each parameter (see Appendix); a mixture of two normal distributions was assumed for the random effect η_2 to reflect the mixture of a diseased population and of a non-diseased one. The results of 500 iterations of this algorithm were kept leading to 500 values of η_2 for each patient.

For the Bayesian parametric method, these values were supposed to follow normal distributions in diseased and non-diseased subjects ($N(\mu_1, \sigma_1^2)$ and $N(\mu_0, \sigma_0^2)$), respectively), whom parameters were to be estimated. Non-informative priors were assumed for these parameters: $p\left(\mu_1, \sigma_1^2\right) \propto \sigma_1^{-2}$ and $p\left(\mu_0, \sigma_0^2\right) \propto \sigma_0^{-2}$. For each iteration of the algorithm, one value of these four parameters was sampled from the respective posterior distribution of each parameter (given in section 2.2) and the corresponding optimal threshold was calculated. The half-range mode (Bickel, 2002), the median, and the mean of the optimal threshold distribution were calculated, as well as the HPD region and quantiles 2.5% and 97.5%. The optimal threshold estimate obtained with the method of Wang and Geisser (2005) was also calculated for comparison (denoted the WG estimate).

Two major parameters were allowed to vary:

- the total number of patients $N = \{100, 200, 400\}$;
- the standard deviation of biomarker distributions in non-diseased and diseased subjects $(\sigma_0, \sigma_1) = \{(0.08, 0.05), (0.07, 0.07), (0.05, 0.08), (0.04, 0.12), (0.03, 0.15)\}$ (cases with larger standard deviation in diseased subjects than in non-diseased ones were favoured because more common).

The NB/NC ratio was fixed to 1 leading to R equal to 1 because the prevalence was 0.5, but for $N = 200$, NB/NC ratios of 0.5 and 2 were also tested.

For the simulations designed to evaluate the Bayesian non-parametric method, R was fixed to one, the standard variations were fixed to 0.07, and different values of N were taken : $\{200, 500, 1000\}$; in this case, the results were compared to those obtained with the Bayesian parametric method.

For each set of parameters, 5000 data sets were simulated. Four criteria were used to evaluate the methods:

- the relative bias: defined as the difference between the expected value of the estimated optimal threshold and its theoretical value and expressed as a percentage of the latter value;
- the coverage probability: determined by the percent of simulations for which the credible interval encompassed the theoretical value;
- the coverage symmetry: this occurs when the probability to find the theoretical value above the upper limit of the credible interval is equal to that of finding it below the lower limit;
- the width of the credible interval: it assesses the precision of the estimation method.

For sparsity, in applying the Bayesian non-parametric method, we only show the relative bias and the coverage probability.

3.2 Results of the simulations

For the simulations associated to the Bayesian parametric method, the relative bias of the optimal threshold estimated using the WG method or using the mode, the mean, or the median of its posterior distribution

224 was less than 1% whatever the method, with very similar results using the WG method and the mode, the
225 mean or the median (Table 1). There was a slight increase in the relative bias when the ratio of the standard
226 deviations in the two populations was far from one.

227 The coverage probability of the 95% credible intervals obtained using the HPD or the quantile meth-
228 ods varied from 94% to 95% in most cases, except when the ratio of the standard deviations in the two
229 populations was far from one in which case the coverage probability was never lower than 92.5%. The
230 credible intervals were generally symmetric, except again when the ratio of the standard deviations in the
231 two populations was far from one; credible intervals obtained using the quantile method were generally a
232 little more asymmetric than the ones obtained using the HPD method.

233 The width of the credible intervals decreased with the number of subjects.

Table 1 Relative bias (RB) 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.

N	NB/NC	σ_0	σ_1	RB WG	RB Mode	RB Median	RB Mean	CP HPD*	1-LCP† HPD*	1-RCP‡ HPD*	WCI HPD*	CP Quant§	1-LCP† Quant§	1-RCP‡ Quant§	WCI Quant§
100	1	0.08	0.05	-0.0013	-0.0013	-0.0033	-0.0025	0.9470	0.0232	0.0298	0.0290	0.9488	0.0204	0.0308	0.0293
100	1	0.07	0.07	-0.0008	-0.0006	-0.0009	-0.0008	0.9380	0.0326	0.0294	0.0309	0.9450	0.0290	0.0260	0.0312
100	1	0.05	0.08	0.0015	0.0009	0.0032	0.0025	0.9448	0.0298	0.0254	0.0296	0.9484	0.0292	0.0224	0.0299
100	1	0.04	0.12	0.0055	0.0051	0.0086	0.0073	0.9468	0.0366	0.0166	0.0292	0.9476	0.0394	0.0130	0.0296
100	1	0.03	0.15	0.0093	0.0083	0.0122	0.0108	0.9354	0.0522	0.0124	0.0275	0.9306	0.0586	0.0108	0.0279
200	0.5	0.08	0.05	-0.0031	-0.0032	-0.0049	-0.0043	0.9398	0.0206	0.0396	0.0225	0.9422	0.0166	0.0412	0.0227
200	0.5	0.07	0.07	-0.0022	-0.0026	-0.0034	-0.0031	0.9418	0.0216	0.0366	0.0233	0.9442	0.0206	0.0352	0.0236
200	0.5	0.05	0.08	-0.0001	-0.0006	-0.0002	-0.0003	0.9396	0.0292	0.0312	0.0205	0.9428	0.0278	0.0294	0.0207
200	0.5	0.04	0.12	0.0035	0.0032	0.0043	0.0040	0.9494	0.0350	0.0156	0.0182	0.9490	0.0374	0.0136	0.0184
200	0.5	0.03	0.15	0.0062	0.0059	0.0073	0.0068	0.9360	0.0542	0.0098	0.0165	0.9378	0.0544	0.0078	0.0167
200	1	0.08	0.05	-0.0017	-0.0016	-0.0027	-0.0023	0.9450	0.0254	0.0296	0.0204	0.9478	0.0232	0.0290	0.0206
200	1	0.07	0.07	-0.0008	-0.0011	-0.0008	-0.0008	0.9424	0.0242	0.0334	0.0218	0.9484	0.0220	0.0296	0.0220
200	1	0.05	0.08	0.0015	0.0014	0.0023	0.0020	0.9416	0.0312	0.0272	0.0208	0.9466	0.0296	0.0238	0.0210
200	1	0.04	0.12	0.0051	0.0049	0.0067	0.0061	0.9436	0.0392	0.0172	0.0205	0.9450	0.0412	0.0138	0.0207
200	1	0.03	0.15	0.0100	0.0095	0.0114	0.0107	0.9262	0.0680	0.0058	0.0193	0.9250	0.0700	0.0050	0.0195
200	2	0.08	0.05	-0.0017	-0.0009	-0.0016	-0.0014	0.9402	0.0274	0.0324	0.0203	0.9440	0.0242	0.0318	0.0205
200	2	0.07	0.07	0.0006	0.0010	0.0021	0.0017	0.9410	0.0314	0.0276	0.0235	0.9446	0.0312	0.0242	0.0237
200	2	0.05	0.08	0.0036	0.0035	0.0054	0.0047	0.9428	0.0370	0.0202	0.0230	0.9446	0.0372	0.0182	0.0233
200	2	0.04	0.12	0.0075	0.0071	0.0096	0.0087	0.9394	0.0450	0.0156	0.0236	0.9398	0.0472	0.0130	0.0239
200	2	0.03	0.15	0.0127	0.0117	0.0143	0.0133	0.9282	0.0668	0.0050	0.0221	0.9260	0.0704	0.0036	0.0223
400	1	0.08	0.05	-0.0022	-0.0021	-0.0027	-0.0024	0.9472	0.0200	0.0328	0.0144	0.9484	0.0188	0.0328	0.0146
400	1	0.07	0.07	-0.0006	-0.0005	-0.0006	-0.0006	0.9434	0.0252	0.0314	0.0180	0.9456	0.0244	0.0300	0.0182
400	1	0.05	0.08	0.0015	0.0015	0.0019	0.0018	0.9448	0.0326	0.0226	0.0147	0.9508	0.0298	0.0194	0.0148
400	1	0.04	0.12	0.0056	0.0056	0.0064	0.0061	0.9320	0.0562	0.0118	0.0144	0.9350	0.0556	0.0094	0.0145
400	1	0.03	0.15	0.0095	0.0092	0.0102	0.0099	0.9242	0.0734	0.0024	0.0135	0.9250	0.0728	0.0022	0.0137

† 1-LCP: one minus left coverage probability; ‡ 1-RCP: one minus right coverage probability; * HPD: credible interval with the highest posterior density method; § Quant: credible interval with the quantiles method.

For the simulations associated to the non-parametric method, the relative bias with that method was always higher than that of the parametric method, but always smaller than 0.2%, which is quite acceptable (Table 2). The coverage probability was close to 95% except with $N = 200$.

Table 2 Relative bias (RB) of the optimal threshold estimate and coverage probability (CP) of the associated credible interval, with the parametric and non-parametric methods and for different sample sizes.

N	Parametric method					Non-parametric method				
	RB Mode	RB Median	RB Mean	CP HPD*	CP Quant§	RB Mode	RB Median	RB Mean	CP HPD*	CP Quant§
1000	-0.0009	-0.0009	-0.0009	0.9348	0.941	0.0041	0.0033	0.0035	0.9688	0.973
500	-0.0003	-0.0003	-0.0003	0.943	0.9464	0.0033	0.0024	0.0027	0.9412	0.9514
200	-0.0001	-0.0002	-0.0002	0.945	0.9452	-0.0005	-0.002	-0.0015	0.8692	0.8862

* HPD: credible interval with the highest posterior density method; § Quant: credible interval with the quantiles method.

4 Example: Diagnosis of Prostate Cancer Recurrence

4.1 Context

The present study involved patients who attended Edouard Herriot Hospital (Lyon, France) between 2000 and 2007 and were offered HIFU as a curative-intent local therapy for prostate cancer. After therapy, the patients were followed with PSA measurements every three weeks during the first four months, every month until the eighth one, and then every four months. For this analysis, we retained only patients with no previous hormonal therapy and at least one biopsy and one PSA measurement after a 90-day follow-up. A more detailed description of the study is given in Subtil and Rabilloud (2010). Among the 289 patients, 150 were declared as developing a recurrence (diseased subjects) from biopsy results. There was

a quick decline of PSA after therapy, potentially followed by an increase. Besides this general pattern, there were some occasional fluctuations possibly due to inherent biological PSA variability or to factors such as infection, prostate manipulation, or measurement errors (Soletormos et al., 2005), but not to progression toward recurrence.

The smallest PSA value reached during the follow-up –the nadir– has been found to be the best dynamic criterion for the diagnosis of prostate cancer recurrence (Subtil and Rabilloud, 2010). The PSA nadir determination directly from the observed measurements may depend on the frequency of measurements and be greatly influenced by occasional PSA fluctuations unrelated to prostate cancer recurrence; thus, to limit bias in nadir estimates, these estimates were calculated from modelled PSA trajectories. An exponential decay-exponential growth mixed model was used to model the PSA trajectories:

$$y_{lj} = \exp(r_{1l}) \exp(-r_{2l} t_{lj}) + \exp(r_{3l}) \exp(r_{4l} t_{lj})$$

where t_{lj} denotes the time elapsed since HIFU treatment and $\{r_1, r_2, r_3, r_4\}$ are random effects. Q-Q plots (not shown) of the predicted random effects rejected the hypothesis of normality for r_1 and r_4; this hypothesis was relaxed by assuming a Dirichlet process as prior for the distributions of these two random effects (Brown and Ibrahim, 2003; Escobar and West, 1998). Moreover, a Student-t distribution was used for residuals to take into account occasional PSA fluctuations. A MCMC algorithm, described in (Subtil and Rabilloud, 2010), was used for inference; 3000 iterations were kept, leading to 3000 samples from the posterior distributions of random effects for each patient. The nadir was easily calculated for each patient at each iteration using random effects values:

$$Nadir_l = \exp(r_{1l}) \exp(-r_{2l} T_l) + \exp(r_{3l}) \exp(r_{4l} T_l), \text{ with } T_l = \frac{1}{r_{2l} + r_{4l}} \log\left(\frac{\exp(r_{1l})\, r_{2l}}{\exp(r_{3l})\, r_{4l}}\right) \quad (2)$$

The nadir values were found to follow lognormal distributions in diseased and non-diseased subjects over all iterations. Non-informative priors were assumed for the parameters of these lognormal distributions; combining them with the nadir values sampled for diseased and non-diseased subjects, 3000 samples from the posterior distributions of the parameters of the nadir distributions were obtained, from which the optimal threshold and its credible interval were calculated as in section 2.3.

The prevalence of the disease was estimated to be 0.52. A sensitivity analysis was carried out for the NB/NC ratio, with values set at $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, \text{ and } 4\}$.

The non-parametric method might not be applicable to these data because there were not enough patients in each group. The optimal threshold and corresponding credible interval obtained with this second method are presented only for $NB/NC = 1$, and should be interpreted with caution.

4.2 Results

Figure 1 represents the posterior distribution of the optimal threshold for $NB/NC = 1$ and for the parametric method, from which a point estimate and a credible interval can be obtained. Depending on the NB/NC values, the optimal threshold estimates ranged from 0.06 to 0.82, with small differences between the mode, the median, and the mean estimates (Table 3). Except for $NB/NC = 0.5$, the credible intervals were very similar with the quantile or the HPD method. The sensitivity relative to the optimal threshold varied from 0.41 to 0.94, whereas the specificity ranged from 0.24 to 0.79.

With the non-parametric method, the estimates of the optimal threshold, the sensitivity, and the specificity were similar to the ones obtained with the parametric method. However the credible intervals were different, probably because of the small number of patients in each group.

In their study about disease-free survival after HIFU treatment, Ganzer et al. (2008) have estimated the hazard ratio of the disease-free survival of a nadir between 0.21 and 1 ng/mL to a nadir lower than 0.2 ng/mL to be 7.436 (1.260, 43.894). They stated that promising oncological outcome could be obtained if a PSA nadir lower than 0.2 ng/mL is reached. Our results are in agreement with those results. The specificity,

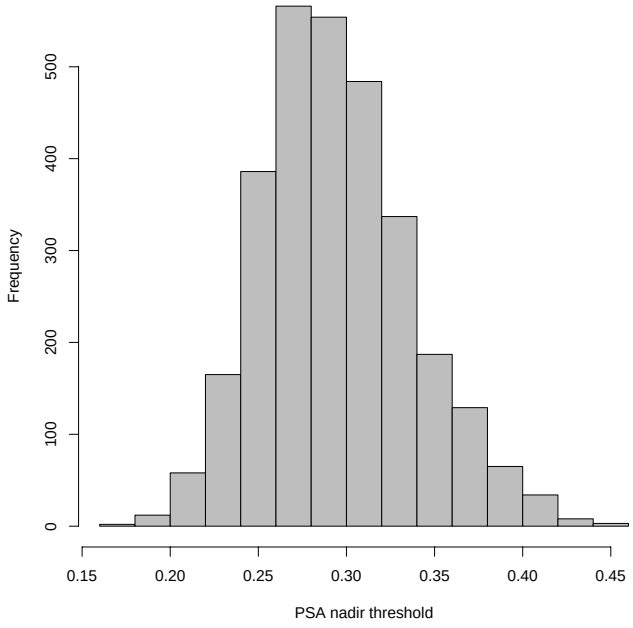

Fig. 1 Posterior distribution of the optimal nadir threshold for $NB/NC = 1$

was quite low, but this is a common feature of PSA whose levels may change just because of an infection. We have not defined a unique threshold but rather a range of values that depend on the NB/NC ratio. A specific threshold value can be estimated using the exact NB/NC value given by a panel of clinicians, but it is perhaps more appropriate to allow this threshold to vary according to different individual choices regarding quality of life. The use of decision curves, proposed by Vickers and Elkin (2006), may help taking into account the uncertainty about the NB/NC value.

Table 3 PSA nadir thresholds estimates and credible intervals (CI) for various values of NB/NC.

Method	NB/NC	Mode	Median	Mean	CI (HPD*)	CI (Quantile)	Sensitivity [CI]	Specificity [CI]
Parametric								
	0.5	0.722	0.795	0.884	[0.450, 1.450]	[0.510, 1.774]	0.411 [0.344, 0.480]	0.788 [0.717, 0.848]
	1	0.273	0.291	0.296	[0.219, 0.388]	[0.221, 0.392]	0.708 [0.642, 0.772]	0.559 [0.478, 0.634]
	1.5	0.184	0.182	0.181	[0.124, 0.239]	[0.122, 0.239]	0.813 [0.754, 0.867]	0.443 [0.359, 0.525]
	2	0.140	0.135	0.133	[0.077, 0.184]	[0.077, 0.185]	0.866 [0.812, 0.912]	0.372 [0.289, 0.456]
	2.5	0.102	0.108	0.107	[0.052, 0.153]	[0.054, 0.155]	0.896 [0.849, 0.936]	0.324 [0.243, 0.408]
	3	0.088	0.091	0.090	[0.039, 0.135]	[0.039, 0.135]	0.916 [0.874, 0.951]	0.288 [0.209, 0.371]
	3.5	0.075	0.079	0.078	[0.028, 0.120]	[0.029, 0.122]	0.930 [0.891, 0.961]	0.261 [0.184, 0.341]
	4	0.065	0.070	0.070	[0.025, 0.113]	[0.024, 0.112]	0.940 [0.904, 0.968]	0.239 [0.163, 0.318]
Non-parametric								
	1	0.212	0.220	0.240	[0.148, 0.326]	[0.164, 0.711]	0.768 [0.736, 0.798]	0.509 [0.464, 0.550]

* HPD: credible interval with the highest posterior density method.

5 Discussion

In the present article, the Bayesian inference has been used to estimate the optimal threshold of a quantitative biomarker and its credible interval. Two methods have been proposed: a parametric one, applicable to fixed and longitudinal biomarkers, and a non-parametric one, applicable to longitudinal biomarkers only. With the parametric method, the results of the numerical study in the case of longitudinal biomarkers underlined that the relative bias of the estimate was very low and comparable to the one obtained using the method of Wang and Geisser (2005). Whatever the parameter settings, this bias was lower to 1%. More interesting was the fact that the coverage probability of the credible interval was close to 0.95 whatever the parameter settings.

In fact, there are only few previous works on the confidence interval of an optimal threshold estimate. Bootstrap methods were used whenever no explicit formula of the optimal threshold was available but, then, in some cases, the coverage probability is likely to be far from 0.95. The HPD and the quantile methods gave similar results, though there was less asymmetry in some cases with the former method. Concerning the choice between the mode, the mean, and the median, our results cannot favour one in particular, even if it can be noted that the mean and the median tended to yield very similar results.

For longitudinal biomarkers, two solutions can be used to estimate the optimal threshold of a diagnostic criterion issued from the serial measurements: the first is to calculate directly the diagnostic criterion from raw measurements, and then use the method described in section 2.2; the second is to model biomarker trajectories and then calculate the diagnostic criterion from modelled trajectories, and estimate the threshold with the method presented in section 2.3. The choice of the method depends on the context but the second method seems more appropriate when the diagnostic criterion is sensitive to the measurement frequency. In the PSA study, the nadir values calculated directly from raw measurements were dependent on the measurement frequency and also biased by occasional PSA fluctuations; this is why the second method was favoured.

The proposed method can be used whatever the distribution family of the marker values (which may be different in diseased and non-diseased subjects) providing the distribution of the marker values is known and the sampling from the posterior distributions of the parameters of the biomarker value distributions possible. The latter condition is met with gamma, normal or lognormal distributions or with normal distributions obtained after a Box-Cox transformation. More research is needed to work with more flexible distributions. If the distribution chosen for the biomarker is far from the true one, for example a mixture of normal distributions instead of only one, the parametric method leads to biased estimates of the optimal threshold, and to coverage probability far from 95% (data not shown). However, in the case of a diagnostic criterion calculated from a model, one can use the non-parametric method if the number of patients is higher than 500 (250 in each group); which yields acceptable coverage probability and relative bias.

The proposed method relies on Bayesian inference. Here, non-informative priors were used, but informative ones may also be used whenever available. Using samples from the posterior distributions of the model parameters allows inferring immediately about other quantities derived from these parameters; this great principle in Bayesian inference allowed us to obtain the posterior distribution of the optimal threshold from the posterior distributions of the parameters of the biomarker distribution. To our knowledge, this is the first time that this method is used to estimate a parameter (and its credible interval), which is the solution of an optimisation problem. Methods proposed in Bayesian optimisation generally focus on the optimisation of the expectation of the utility function (or optimisation function) over the distribution of the parameters, and in most cases, does not provide credible intervals (Chaloner and Verdinelli, 1995).

The literature dedicated to the optimal threshold has proposed methods to deal with measurement errors (Perkins and Schisterman, 2005), marker with mass at zero (Schisterman et al., 2008), or detection threshold (Ruopp et al., 2008). Our method can be extended to these cases by taking into account these features in the inference on the parameters of the biomarker distributions. The method can be also extended to the case of optimal dichotomization for repeated screening tests, using the work of Wang and Geisser (2003), with the advantage of providing a simple way to construct a credible interval.

Concerning the application part, depending on the NB/NC value, the optimal threshold ranged from 0.07 to 0.82, with credible interval widths ranging from 0.04 to 0.5. Our results will help clinicians monitor patients after HIFU treatment; it is indeed, to our knowledge, the first time that a threshold value is defined for the nadir in terms of optimal utility.

Acknowledgements This study was partially funded by the French "Ligue contre le Cancer". The authors are grateful to Dr. Albert Gelet of the "Service d'Urologie et de Chirurgie de la Transplantation" of the "Hospices Civils de Lyon" for providing the data of PSA after HIFU. We also thank Jean Iwaz for suggestions and criticisms concerning the manuscript.

6 Appendix

6.1 Derivation of the PSA nadir optimal threshold

The derivation of the PSA nadir optimal threshold was obtained in three parts:

- first, sample from the posterior distribution of the nadir for each patient (1);
- second, sample from the posterior distribution of the parameters of the distribution of the nadir in the diseased and non-diseased group (2);
- third, construct the posterior distribution of the PSA nadir threshold from the posterior distribution of the utility function (3).

(1) A Gibbs sampler algorithm with K iterations, described in Subtil and Rabilloud (2010), was used to sample K values from the posterior distributions of the random effects for each patient $\{r_{1l}, r_{2l}, r_{3l}, r_{4l}\}$. For the l^{th} patient, the values of r_{1l}, r_{2l}, r_{3l} and r_{4l} at the k^{th} iteration of the Gibbs sampler were combined using equation 2 to obtain an estimate of the nadir for this patient. The values obtained from the K iterations forms the posterior distribution of the nadir for patient l. This process was repeated for each patient.

(2) Afterwards, at each iteration, it was found that the nadir estimates in the non-diseased patients were lognormally distributed, and it was also the case in the diseased patients. Let μ_{0k} and σ^2_{0k} (resp. μ_{1k} and σ^2_{1k}) denotes the mean and variance parameter of distribution of the logarithm of the nadir at iteration k in the non-diseased patients (resp. in the diseased patients). Under the non-informative prior $p(\mu_{0k}, \sigma^2_{0k}) \sim \sigma^{-2}_{0k}$, the marginal posterior distribution of σ^2_{0k} was:

$$\left(\log(\mathbf{Nadir})_{0k} - \overline{\log(\mathbf{Nadir})_{0k}}\right)^T \left(\log(\mathbf{Nadir})_{0k} - \overline{\log(\mathbf{Nadir})_{0l}}\right) / \chi^2_{n_0-1}$$

where $\log(\mathbf{Nadir})_{0k}$ denotes the vector of the logarithm of the nadir estimates in non-diseased patients at iteration k. The conditional distribution of μ_{0k} given σ_{0k} was normal: $N(\overline{\mathbf{Nadir}_{0k}}, \sigma^2_{0k}/n_0)$. At each iteration, one could sample a value of σ_0 and μ_0 from their posterior distribution. The same was done for σ_1 and μ_1.

(3) From these estimates, at each iteration, a utility function was build, with the sensitivity and specificity obtained from the repartition function of the normal distribution of the logarithm of the nadir in the diseased and non-diseased patients at the specific iteration ($N(\mu_{1k}, \sigma^2_{1k})$ and $N(\mu_{0k}, \sigma^2_{0k})$). By maximising these utility functions, each iteration gave a value of the optimal nadir threshold; their values over all iterations form the posterior distribution of the logarithm of the optimal nadir threshold.

6.2 Model used for the simulation part

The general model for the estimation of the random effects in simulations was:

$$\log(y_{lj}) = \log\left(\exp(\eta_{1l}) + \exp(\eta_{2l} t_{lj})\right) + \varepsilon_{lj}$$

$$\varepsilon_{lj} \sim N\left(0, \sigma^2_\varepsilon\right) \quad \eta_{1l} \sim N\left(\mu_{\eta_1}, \sigma^2_{\eta_1}\right) \quad \eta_{2l} \sim 0.5 \times N\left(\mu_{\eta_2}, \sigma^2_{\eta_{20}}\right) + 0.5 \times N\left(\mu_{\eta_2} + \delta, \sigma^2_{\eta_{21}}\right)$$

Explicit forms of the full conditional posterior distributions existed for all parameters, except for random effects (due to the non-linearity of the model). They were consequently sampled using a Metropolis algorithm with normal proposal; the proposal was adjusted to obtain an acceptance rate close to 0.44, as

suggested by Gelman et al. (2004). Non-informative normal prior distributions were used for μ_{η_1} and μ_{η_2} ($N(0, 1/0.001)$). A non-informative normal distribution truncated for values below 0 was used for δ ($N(0, 1/0.001)I(0,)$) to ensure the identifiability of the model –i.e. diseased patients had a lower increase than non-diseased ones– and non-informative Gamma prior distributions were used for $1/\sigma^2_{\varepsilon}$, $1/\sigma^2_{\eta_1}$, $1/\sigma^2_{\eta_{20}}$ and $1/\sigma^2_{\eta_{21}}$ ($\mathrm{Gamma}(0.001, 0.001)$). Because the likelihood largely dominated prior distributions, the results were not sensitive to prior distributions. The results were based on 1 in 4 thinning on the total of 2000 iterations, after a burn-in period of 4000 iterations. Convergence was checked using multiple chains with overdispersed starting values, and the Gelman and Rubin diagnostic (Gelman and Rubin, 1992). The algorithm was implemented in R (R Development Core Team, 2008). The full conditional posterior densities are given as supplementary material.

References

Bickel, D. R. (2002). Robust estimators of the mode and skewness of continuous data. *Computational Statistics & Data Analysis* **39**, 153–163.

Brown, E. R. and Ibrahim, J. G. (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science* **10**, 273–304.

Chen, L. S., Geisser, S. and Geyer, C. J. (1999). Monte carlo minimization for sequential control. In: *Diagnosis and Prediction* (eds. S. Geisser), Springer-Verlag, New York, 109–130.

DeNeef, P. and Kent, D. L. (1993). Using treatment-tradeoff preferences to select diagnostic strategies: linking the roc curve to threshold analysis. *Medical Decision Making* **13**, 126–32.

405 Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In: *Practical Non-*
406 *parametric and Semiparametric Bayesian Statistics* (eds. D. Dey, P. Müller and D. Sinha), Springer-
407 Verlag, New-York, 1–22.

408 Fluss, R., Faraggi, D. and Reiser, B. (2005). Estimation of the youden index and its associated cutoff point.
409 *Biometrical Journal* **47**, 458–472.

410 Ganzer, R., Rogenhofer, S., Walter, B., Lunz, J.-C., Schostak, M., Wieland, W. F. and Blana, A. (2008).
411 PSA nadir is a significant predictor of treatment failure after high-intensity focussed ultrasound
412 (HIFU) treatment of localised prostate cancer. *European Urology* **53**, 547–553.

413 Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman &
414 Hall/CRC, London.

415 Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical*
416 *Science* **7**, 457–511.

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

419 Laird, N. M., Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

420 Leeflang, M. M. G., Moons, K. G. M., Reitsma, J. B. and Zwinderman, A. H. (2008). Bias in sensitivity
421 and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and
422 solutions. *Clinical Chemistry* **54**, 729–737.

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

425 Perkins, N. J. and Schisterman, E. F. (2005). The youden index and the optimal cut-point corrected for
426 measurement error. *Biometrical Journal* **47**, 428–441.

427 R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Founda-
428 tion for Statistical Computing, Vienna.

Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. and Schisterman, E. F. (2008). Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal* **50**, 419–430.

Schiffer, E. (2009). The 2nd annual oncology biomarkers conference. *Biomarkers in Medicine* **3**, 203–209.

Schisterman, E. F., Faraggi, D., Reiser, B. and Hu, J. (2008). Youden index and the optimal threshold for markers with mass at zero. *Statistics in Medicine* **27**, 297–315.

Schisterman, E. F. and Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics: Simulation and Computation* **36**, 549–563.

Soletormos, G., Semjonow, A., Sibley, P. E., Lamerz, R., Petersen, P. H., Albrecht, W., Bialk, P., Gion, M., Junker, F., Schmid, H.-P. and Van Poppel, H., on behalf of the European Group on Tumor Markers, (2005). Biological variation of total prostate-specific antigen: A survey of published estimates and consequences for clinical practice. *Clinical Chemistry* **51**, 1342–1351.

Sox, H. C., Blatt, M. A., Higgins, M. C. and Marton, K. I. (1988). *Medical Decision Making*. Butterworths, Boston.

Subtil, F., Pouteil-Noble, C., Toussaint, S., Villar, E. and Rabilloud, M. (2009). A simple modeling-free method provides accurate estimates of sensitivity and specificity of longitudinal disease biomarkers. *Methods of Information in Medicine* **48**, 299–305.

Subtil, F. and Rabilloud, M. (2010). Robust non-linear mixed modelling of longitudinal psa levels after prostate cancer treatment. *Statistics in Medicine* **29**, 573–587.

Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer, New York.

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**, 565–574.

Wang, M.-D. and Geisser, S. (2003). Optimal dichotomization for repeated screening tests. *Statistics and Probability Letters* **62**, 61–70.

453 Wang, M.-D. and Geisser, S. (2005). Optimal dichotomization of screening test variables. *Journal of Sta-*
454 *tistical Planning and Inference* **131**, 191–206.

455 Zheng, Y. and Heagerty, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics* **63**, 332–
456 341.

5.4.2 Principaux résultats de l'article

Les résultats des simulations montrent que l'estimation ponctuelle du seuil optimal par la méthode Bayésienne est très peu biaisée, le faible biais subsistant étant similaire en retenant le mode, la médiane ou la moyenne de la distribution a posteriori. Ces estimations ponctuelles sont similaires à celles obtenues par la méthode de Wang et Geisser. Concernant les intervalles de crédibilité à 95 %, la probabilité de couverture des intervalles fournis par la méthode HDP et la méthode des quantiles de la distribution a priori sont souvent très proches et tournent autour de 95 % ; l'une ou l'autre des méthodes peut être utilisée, sans préférence.

L'article présente succinctement les résultats concernant le seuil optimal du nadir de PSA. Ces résultats sont discutés et affinés dans le chapitre suivant.

5.5 Compléments à l'article

5.5.1 Les valeurs limites de BN/CN

L'application de la méthode paramétrique d'estimation du seuil optimal aux données de PSA a fourni des résultats similaires en termes de mode, médiane et moyenne de la distribution a posteriori du seuil, ainsi qu'en termes d'intervalles de crédibilité obtenus par la méthode des quantiles et HDP, sauf pour $BN/CN = 0,5$. Il est intéressant d'observer les courbes de densité de probabilité des valeurs de logarithme de nadir, la courbe des malades étant multipliée par $\pi \times BN$ et celle des non malades par $(1-\pi) \times CN$, et ce, pour les différentes valeurs de BN/CN étudiées dans l'article.

Les graphiques de la figure 5.8 ont été construits en retenant les valeurs de paramètres de distribution du logarithme du nadir chez les malades et les non malades à une itération de l'algorithme MCMC. Pour $BN/CN = 0,5$, les deux courbes sont quasiment emboîtées. Lorsque les courbes sont emboîtées, il n'existe plus de valeur de seuil optimal. Les coûts de la réalisation à tort d'une biopsie sont jugés tels par rapport aux distribution du logarithme du nadir chez les malades et les non malades qu'il vaut mieux ne jamais réaliser de biopsie. Pour $BN/CN = 0,5$, les fluctuations des valeurs de paramètres de distribution du nadir d'une itération à l'autre de l'algorithme MCMC peuvent entraîner la disparition du seuil optimal, d'où une instabilité dans son estimation, qui se traduit par des résultats un peu divergents entre les différentes méthodes d'estimation ponctuelle et d'intervalle de crédibilité. Pour $BN/CN = 3,5$ ou $BN/CN = 4$, les courbes semblent également emboîtées ; en réalité, l'échelle est très différente de celle utilisée

pour $BN/CN = 0,5$. Ainsi, ces courbes ne sont pas emboîtées, et il n'y a pas d'instabilité dans l'estimation du seuil optimal.

Il faut garder à l'esprit qu'un marqueur n'a d'utilité uniquement pour certaines plages de valeurs de BN/CN lorsque l'objectif est de prendre une décision, telle la réalisation d'une biopsie ou d'un traitement ; cette notion sera reprise au cours du chapitre six.

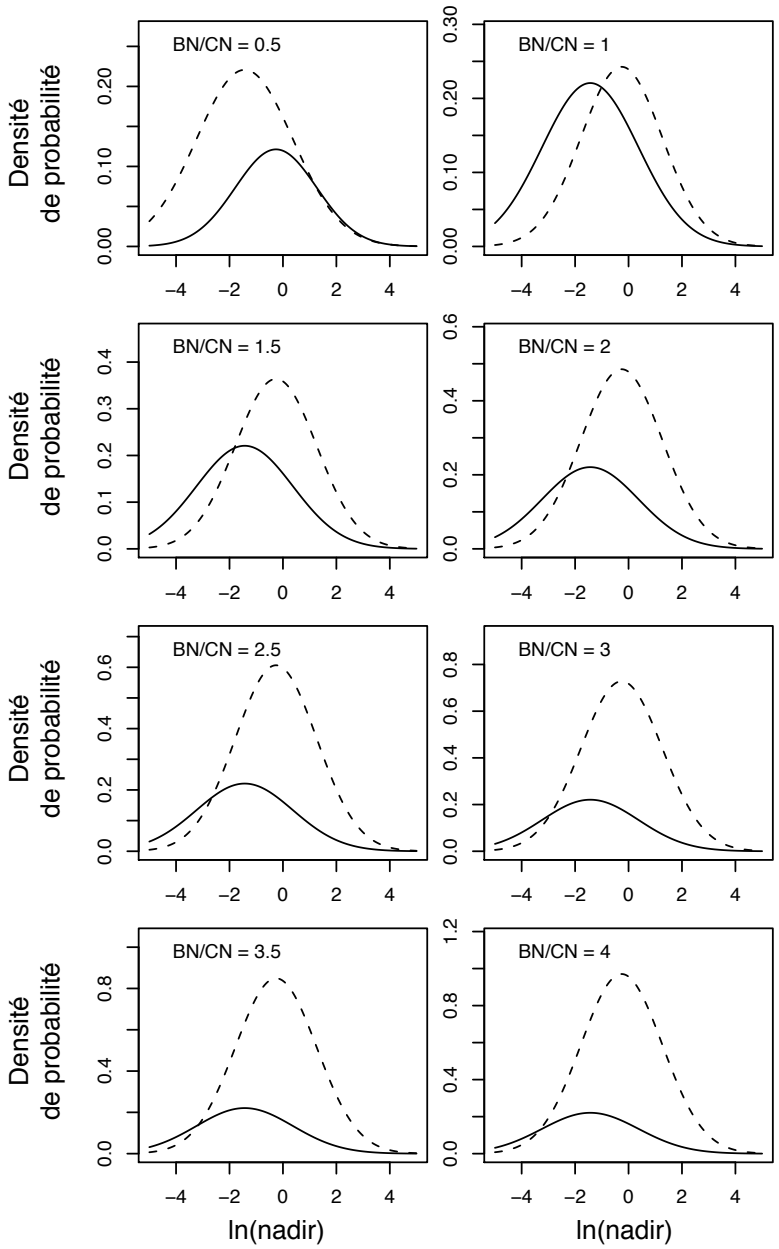

Figure 5.8 – Courbes de densité de probabilité des valeurs de logarithme de nadir, multipliée par $\pi \times BN$ pour les malades (trait plein) et par $(1 - \pi) \times CN$ pour les non malades (trait pointillé), pour différentes valeurs de BN/CN.

5.5.2 Choix de la distribution du marqueur

5.5.2.1 Plus de souplesse pour la modélisation de la distribution du marqueur

La méthode Bayésienne d'estimation du seuil optimal dépend de la loi choisie pour la distribution des valeurs de marqueur dans les deux groupes. Des lois suffisamment souples sont nécessaires pour pouvoir s'adapter facilement à de nombreux marqueurs. Un exemple de loi souple est la loi gamma généralisée (Lawless, 1996), dont l'intérêt dans le cadre de l'estimation du seuil optimal d'un marqueur a été analysé dans ce travail.

La loi gamma généralisée a été introduite pour combiner la puissance de deux distributions : la distribution gamma et la distribution de Weibull. Elle est caractérisée par trois paramètres : a, b et c et sa densité de probabilité en un point $x > 0$ est donnée par :

$$P(x) = \frac{c}{\Gamma(a)} \frac{x^{ca-1}}{b^{ca}} \exp\left(-\left(\frac{x}{b}\right)^c\right)$$

Plusieurs lois sont des sous-cas de la loi gamma généralisée :

- la loi exponentielle est obtenue en fixant $a = c = 1$;
- lorsque $a = 1$, la loi gamma généralisée se simplifie en une loi de Weibull ;
- la loi gamma est obtenue en fixant $c = 1$;
- enfin, la loi log normale est un cas limite de la loi gamma généralisée, lorsque a tend vers l'infini.

Cette loi permet donc de représenter une grande variété de marqueurs ; mais plus une loi est souple, plus ses paramètres sont difficiles à estimer. Lawless (1996) a proposé une autre paramétrisation de la loi, limitant la corrélation entre les paramètres. Elle consiste à prendre : $y = \ln(x)$, $\lambda = 1/\sqrt{a}$, $\sigma = 1/(c\sqrt{a})$ et $\mu = \ln(b) + \ln(a)/c$; la densité de probabilité est alors donnée par :

$$P(y) = \frac{|\lambda|}{\sigma\Gamma(\lambda^{-2})} \exp\left(\lambda^{-2}\left(\lambda\frac{y-\mu}{\sigma} - \exp(\lambda(y-\mu)/\sigma)\right)\right) \tag{5.7}$$

Dans le cadre de l'estimation du seuil d'un marqueur, un a priori non informatif de Jeffrey a été utilisé pour la distribution des trois paramètres de cette loi (Jeffreys, 1946). Cet a priori est proportionnel à la racine carrée du déterminant de la matrice d'information de Fisher ; l'a priori de Jeffrey est donc un a priori non informatif relativement à la quantité d'information contenue dans les données. Dans le cas de la paramétrisation décrite en (5.7), l'a priori de Jeffrey est donné par :

$$-2\frac{\sqrt{(\lambda^{-2}\psi'(|\lambda|^{-2}))^2 - \psi'(|\lambda|^{-2}) - 1}}{\lambda^2\sigma^2}$$

où ψ' est la dérivée de la fonction digammma. La log densité a posteriori des paramètres est donc donnée par :

$$\ln(P(\lambda,\sigma,\mu)|\mathbf{y}) \propto n((1-2\lambda^{-2})\ln(|\lambda|)-\ln(\sigma)-\ln(\Gamma(\lambda^{-2})))+\sum_{i=1}^{n}\lambda^{-2}(\lambda(y_i-\mu)/\sigma-\exp(\lambda(y_i-\mu)/\sigma))$$

Ceci ne correspond pas à une loi de probabilité connue. Un algorithme de type Metropolis multivarié a donc été utilisé pour échantillonner les valeurs des paramètres dans leur distribution a posteriori, avec une loi normale multivariée comme distribution de proposition, centrée sur les valeurs de paramètres de l'itération précédente. La matrice de variance covariance a été fixée dans un premier temps à $(2.4^2 \times \mathbf{\Sigma}/2)$, où $\mathbf{\Sigma}$ est la matrice de variance covariance des estimateurs du maximum de vraisemblance des paramètres. La covariance entre les paramètres a été calculée sur les premières itérations, puis utilisée par la suite comme matrice de variance covariance pour la distribution de proposition, en s'assurant d'obtenir un taux d'acceptation des paramètres proche de 0,23. Cette méthode d'échantillonnage a été proposée par Gelman et *al.* (2004) et donne de bons résultats pour la loi gamma généralisée.

Une fois un échantillon de valeurs des paramètres de la loi gamma généralisée obtenu chez les malades et les non malades, la fonction d'utilité a été calculée pour chaque itération de la chaîne MCMC ; un algorithme de type Newton-Raphson a permis le calcul du seuil maximisant la fonction de l'itération en question. Les seuils obtenus sur l'ensemble des itérations ont constitué la distribution a posteriori du seuil optimal.

Cette méthode d'estimation du seuil et de son intervalle de confiance, dans le cas de la loi gamma généralisée, a été évaluée par simulations, en considérant cinq cas de figures (tableaux 5.1 à 5.3). Dans le premier cas, le marqueur suivait des lois exponentielles chez les malades et les non malades. Pour le deuxième et le troisième cas, le marqueur suivait des lois de Weibull ; pour les deux derniers cas, il suivait des lois gamma. Une loi gamma généralisée a été utilisée pour caractériser la distribution des marqueurs dans chacun des cas. 2000 valeurs de paramètres de la loi ont été échantillonnées dans leur distribution a posteriori via la méthode MCMC proposée précédemment, permettant ainsi l'échantillonnage de valeurs dans la distribution a posteriori du seuil optimal. Pour chaque type de distribution de marqueur, deux paramètres ont été modifiés :

- le nombre de patients dans chacun des groupes ($n = \{50; 100; 200\}$) ;
- le ratio bénéfice net sur coût net ($BN/CN = \{0,5\ ;\ 1;\ 2\}$).

Pour chaque jeu de paramètres de simulation, 5000 jeux de données ont été simulés, permettant le calcul du biais relatif lié à l'estimation ponctuelle du seuil optimal et de la probabilité de couverture de l'intervalle de crédibilité à 95 % (tableaux 5.1 à 5.3).

Tableau 5.1 – Résultats des simulations pour la loi gamma généralisée.

Non malades	Malades	BN/CN	n	Biais relatif			Probabilité de couverture	
				Moyenne	Mode	Médiane	Quant	HDP
$\lambda = 1; \sigma = 1; \mu = -0,69$	$\lambda = 1; \sigma = 1; \mu = 0,69$	0,5	50	0,048	0,017	0,035	0,940	0,943
			100	0,023	0,005	0,016	0,947	0,946
			200	0,01	0,001	0,007	0,952	0,952
		1	50	0,109	0,075	0,096	0,939	0,939
			100	0,041	0,024	0,034	0,951	0,951
			200	0,017	0,009	0,014	0,949	0,950
		2	50	0,061	0,022	0,059	0,971	0,960
			100	0,004	0,01	0,012	0,960	0,952
			200	-0,001	0,008	0,004	0,957	0,956
$\lambda = 1; \sigma = 0,5; \mu = 0$	$\lambda = 1; \sigma = 0,5; \mu = 0,69$	0,5	50	0,025	0,015	0,02	0,939	0,938
			100	0,01	0,004	0,008	0,944	0,945
			200	0,005	0,002	0,004	0,950	0,950
		1	50	0,047	0,042	0,045	0,937	0,931
			100	0,018	0,015	0,017	0,951	0,945
			200	0,007	0,006	0,007	0,951	0,949
		2	50	-0,03	0,022	-0,001	0,961	0,946
			100	-0,02	0,014	-0,002	0,960	0,954
			200	-0,009	0,005	-0,003	0,956	0,957

Tableau 5.2 – Résultats des simulations pour la loi gamma généralisée (suite).

Non malades	Malades	BN/CN	n	Biais relatif			Probabilité de couverture	
				Moyenne	Mode	Médiane	Quant	HDP
$\lambda = 1; \sigma = 0,33; \mu = 0$	$\lambda = 1; \sigma = 0,33; \mu = 0,69$	0,5	50	0,018	0,013	0,016	0,911	0,913
			100	0,011	0,008	0,010	0,923	0,922
			200	0,007	0,006	0,007	0,939	0,940
		1	50	0,032	0,031	0,031	0,913	0,907
			100	0,018	0,017	0,017	0,932	0,927
			200	0,011	0,011	0,011	0,940	0,939
		2	50	0,033	0,040	0,036	0,935	0,920
			100	0,019	0,021	0,020	0,943	0,938
			200	0,012	0,013	0,012	0,944	0,941
$\lambda = 0,7; \sigma = 0,7; \mu = 0$	$\lambda = 0,7; \sigma = 0,7; \mu = 0,69$	0,5	50	0,027	-0,005	0,009	0,942	0,944
			100	0,027	0,005	0,017	0,957	0,962
			200	0,017	0,007	0,013	0,950	0,954
		1	50	0,131	0,112	0,124	0,912	0,909
			100	0,064	0,054	0,060	0,930	0,929
			200	0,035	0,029	0,033	0,947	0,947
		2	50	0,047	0,065	0,077	0,968	0,948
			100	-0,046	0,006	-0,011	0,967	0,946
			200	-0,040	0,016	-0,010	0,965	0,952

Tableau 5.3 – Résultats des simulations pour la loi gamma généralisée (fin).

Non malades	Malades	BN/CN	n	Biais relatif			Probabilité de couverture	
				Moyenne	Mode	Médiane	Quant	HDP
$\lambda = 0,81; \sigma = 0,81; \mu = 0,40$	$\lambda = 0,58; \sigma = 0,58; \mu = 1,1$	0,5	50	0,053	0,018	0,034	0,947	0,948
			100	0,041	0,012	0,027	0,949	0,956
			200	0,028	0,011	0,021	0,956	0,962
		1	50	0,167	0,141	0,156	0,886	0,897
			100	0,082	0,070	0,077	0,922	0,926
			200	0,041	0,035	0,039	0,939	0,941
		2	50	0,118	0,130	0,125	0,949	0,938
			100	0,045	0,051	0,048	0,959	0,955
			200	0,015	0,018	0,017	0,960	0,957

A l'exception d'un jeu de paramètres, les probabilités de couverture sont proches de 95 % et similaires pour la méthode HDP et la méthode des quantiles. Le biais relatif lié à l'estimation du seuil optimal diminue plus la taille de l'échantillon augmente. Pour $n = 50$, le biais relatif maximum est de 18 %, mais il passe à 8 % dès que le nombre de patients dans chaque groupe atteint 100. Les estimations obtenues avec la moyenne, la médiane ou le mode de la distribution a posteriori du seuil sont similaires en termes de biais relatif.

L'utilisation de la loi gamma généralisée et de l'algorithme proposé pour échantillonner dans la distribution a posteriori des paramètres de la loi semble donc fournir de bons résultats concernant l'estimation du seuil optimal et de son intervalle de crédibilité. Le fait que la loi gamma généralisée soit relativement souple pourrait permettre d'amoindrir le problème de non robustesse de la méthode Bayésienne quant au choix de la distribution du marqueur.

5.5.2.2 La loi des valeurs extrêmes dans le cas des données de PSA après UFHI

Le nadir correspondant à la valeur minimale de PSA atteinte par un patient au cours de son suivi, la distribution dite des valeurs extrêmes peut également être envisagée (Gumbel, 1966). Cette loi de probabilité est caractérisée par trois paramètres, μ, σ et ξ ; la densité de probabilité est donnée par :

$$P(y) = \frac{1}{\sigma}\left(1 + \xi\frac{(y-\mu)}{\sigma}\right)^{-1/\xi-1} \exp\left(-\left(1 + \xi\frac{(y-\mu)}{\sigma}\right)^{-1/\xi}\right)$$

Un test d'adéquation de Kolmogorov-Smirnov a été réalisé à chaque itération sur l'ensemble des valeurs de nadir échantillonnées chez les malades et les non malades ; il a montré une adéquation de la loi des valeurs extrêmes avec les nadirs de PSA dans les deux groupes. Ainsi, la loi log normale n'était pas l'unique loi possible pour décrire la distribution des nadirs de PSA. L'objectif de cette partie est d'estimer le seuil optimal obtenu en utilisant une distribution des valeurs extrêmes, puis de comparer les résultats à ceux obtenus avec la distribution log normale.

Des a priori non informatifs ont été utilisés pour les paramètres de la distribution des valeurs extrêmes, l'échantillonnage des valeurs de paramètres étant réalisé par une technique similaire à celle utilisée dans le cas de la loi gamma généralisée. Les résultats, en termes d'estimation ponctuelle du seuil et d'intervalle de crédibilité, sont présentés dans le tableau 5.4, ainsi que ceux concernant la loi log normale. Contrairement à l'article accepté dans le Biometrical Journal, les préférences des patients et des cliniciens n'ont pas été caractérisées en termes de

ratio bénéfice net sur coût net, mais en termes de risque de biopsie positive à partir duquel un patient accepte la réalisation d'une biopsie (r), sachant que $(1-r)/r = BN/CN$.

Tableau 5.4 – Comparaison des estimations de seuil optimal de nadir de PSA obtenues avec la loi log normale et la loi des valeurs extrêmes.

r	Loi	Mode	Médiane	Moyenne	IC* Quant.	IC* HDP
0,6	Log normale	0,477	0,506	0,529	[0,360 ; 0,824]	[0,323 ; 0,747]
	Valeurs extrêmes	0,378	0,406	0,430	[0,287 ; 0,733]	[0,260 ; 0,633]
0,5	Log normale	0,273	0,291	0,296	[0,221 ; 0,392]	[0,219 ; 0,388]
	Valeurs extrêmes	0,230	0,241	0,245	[0,173 ; 0,337]	[0,166 ; 0,324]
0,4	Log normale	0,184	0,182	0,181	[0,122 ; 0,239]	[0,124 ; 0,239]
	Valeurs extrêmes	0,150	0,157	0,158	[0,105 ; 0,219]	[0,101 ; 0,213]
0,3	Log normale	0,111	0,115	0,114	[0,060 ; 0,164]	[0,059 ; 0,163]
	Valeurs extrêmes	0,100	0,099	0,094	[-0,023 ; 0,146]	[-0,02 ; 0,148]
0,2	Log normale	0,065	0,070	0,070	[0,024 ; 0,112]	[0,025 ; 0,113]
	Valeurs extrêmes	0,066	0,048	0,030	[-0,055 ; 0,090]	[-0,052 ; 0,091]

* IC : intervalle de crédibilité

Les estimations ponctuelles obtenues avec les deux lois sont relativement proches, à l'exception de celles obtenues pour $r = 0,6$. Les seuils obtenus avec la loi des valeurs extrêmes sont un peu plus faibles que ceux obtenus avec la loi log normale. Les intervalles de crédibilité sont un peu plus resserrés avec la loi des valeurs extrêmes. Dans certains cas, la borne gauche de l'intervalle de crédibilité obtenu avec cette dernière méthode est négative, ce qui montre la limite de la distribution des valeurs extrêmes dans ce cas, car elle ne contraint pas les valeurs à être positives.

Dans l'ensemble, les résultats obtenus avec les deux méthodes sont assez semblables. Ainsi, même si la méthode Bayésienne d'estimation du seuil est sensible aux choix de la distribution des valeurs de marqueurs, on peut être confiant dans les résultats obtenus, puisqu'ils sont similaires avec deux types de lois différentes pour la modélisation de la distribution des nadirs de PSA dans les deux groupes.

5.5.3 Estimation non paramétrique du seuil optimal

5.5.3.1 Entre espérer et moyenner

L'article accepté dans le Biometrical Journal présente une approche non paramétrique d'estimation du seuil optimal dans le cas d'un biomarqueur dynamique. Le principe consiste, à chaque itération, à calculer la fonction d'utilité grâce aux estimations empiriques de sensibilité et de spécificité basées sur les valeurs de nadirs échantillonnées. Les résultats des simulations montrent que la probabilité de couverture des intervalles de crédibilité fournis par cette méthode n'est acceptable que lorsque les estimations sont basées sur un très grand nombre d'individus. Pour comprendre ce résultat, il est nécessaire de formaliser cette méthode.

L'estimation du seuil optimal repose sur le calcul de l'utilité espérée par la méthode prédictive (partie 5.1.4.2). Cette utilité s'écrit :

$$\begin{aligned}\mathcal{U}^{\star}(c) &= P(\widetilde{\text{nadir}}_1 > c|\mathbf{y}) + P(\widetilde{\text{nadir}}_0 \leq c|\mathbf{y})R \\ &= \int\int \Big(P(\widetilde{\text{nadir}}_1 > c|\mathbf{nadir}_1) + P(\widetilde{\text{nadir}}_0 \leq c|\mathbf{nadir}_0)R\Big) \\ &\qquad \times P(\mathbf{nadir}_1, \mathbf{nadir}_0|\mathbf{y})\; d\mathbf{nadir}_1 d\mathbf{nadir}_1 \end{aligned} \tag{5.8}$$

$\widetilde{\text{nadir}}_1$ correspond à la valeur prédite de nadir de PSA pour un malade et $\widetilde{\text{nadir}}_0$ à celle prédite pour un non malade. $P(\mathbf{nadir}_1, \mathbf{nadir}_0|\mathbf{y})$ dénote la probabilité a posteriori des n_1 valeurs de nadirs chez les malades et des n_0 valeurs chez les non malades, d'après les mesures $\mathbf{y}$ de PSA. La sensibilité et la spécificité ont ensuite été estimées de façon empirique, à partir des échantillons de valeurs de nadirs pour chaque patient dans les deux groupes :

$$\begin{aligned}\mathcal{U}^{\star}(c) = \int\int \left(\sum_{i=1}^{n_1} I(\text{nadir}_{1i} > c)/n_1 + \sum_{i=1}^{n_0} I(\text{nadir}_{0i} \leq c) \times R/n_0\right) \\ \times P(\mathbf{nadir}_1, \mathbf{nadir}_0|\mathbf{y})\; d\mathbf{nadir}_1 d\mathbf{nadir}_1\end{aligned} \tag{5.9}$$

L'intégrale précédente a été calculée par la méthode de Monte Carlo, grâce aux valeurs de nadirs échantillonnées au travers de la chaîne MCMC. Chaque itération de cette chaîne a conduit à une fonction d'utilité, donc à la détermination d'un seuil optimal ; les seuils obtenus sur l'ensemble des itérations ont formé la distribution a posteriori du seuil optimal.

Dans l'équation (5.9), l'utilité espérée, obtenue à partir des prédictions a posteriori de la fonction d'utilité pour de nouveaux patients, a été remplacée par une utilité moyenne sur les patients de l'étude. Par conséquent, les valeurs échantillonnées de seuil optimal ne correspondent

plus à la distribution a posteriori du seuil maximisant l'utilité espérée, expliquant pourquoi la probabilité de couverture de l'intervalle de crédibilité n'était pas toujours acceptable.

En passant de l'utilité espérée à l'utilité moyenne, une partie de la variabilité des prédictions est perdue. Dans l'équation (5.8), une estimation de la sensibilité prédite a posteriori ($P(\widetilde{\text{nadir}}_1 > c|\mathbf{nadir}_1)$) peut être obtenue en tirant une valeur dans une loi binomiale de paramètres n_1 et $\sum_{i=1}^{n_1} I(\text{nadir}_{1i} > c)/n_1$, puis en divisant cette valeur par n_1. Dans le cas de l'équation (5.9), qui est l'approche proposée dans l'article accepté dans le Biometrical Journal, cette sensibilité est remplacée directement par $\sum_{i=1}^{n_1} I(\text{nadir}_{1i} > c)/n_1$. Or les échantillons issus d'une loi binomiale sont plus variables que l'espérance de cette loi binomiale, sauf lorsque le nombre d'individus devient important. C'est pour cela que, lorsque le nombre de patients est élevé, la probabilité de couverture de l'intervalle de crédibilité obtenu par la méthode non paramétrique d'estimation du seuil est acceptable.

Le fait de remplacer des prédictions a posteriori par des estimations empiriques avait déjà été proposé par Scarpa et Dunson (2007), dans un problème d'optimisation différent de celui de l'estimation du seuil optimal d'un marqueur. Néanmoins, les limites de cette approche n'avaient pas été discutées par les auteurs. Elles sont bien visibles dans le cas de l'estimation du seuil optimal d'un marqueur et de son intervalle de crédibilité, où il est rare de disposer de mesures chez un très grand nombre de patients.

5.5.3.2 Approche semi-paramétrique prédictive

⊂ Principe

Pour déterminer la probabilité a posteriori qu'une valeur de marqueur soit supérieure à un seuil donné sans remplacer cette probabilité par une estimation empirique, il est nécessaire de connaître la distribution du marqueur dans le groupe considéré. Néanmoins, il existe des méthodes non paramétriques pour la modélisation d'une distribution de probabilité, comme les processus de Dirichlet introduits dans le chapitre quatre. A chaque itération, la distribution des valeurs de marqueurs chez les malades et les non malades peut être modélisée grâce à un processus de Dirichlet ; une fois les paramètres du processus de Dirichlet échantillonnés, il est possible de déterminer la probabilité a posteriori qu'une valeur de marqueur soit supérieure ou inférieure à un seuil. Dans ce cas, les fonctions d'utilité construites correspondent bien à des prédictions a posteriori de l'utilité ; les valeurs de seuil optimal échantillonnées constituent bien la distribution a posteriori du seuil optimal maximisant l'utilité espérée.

Les processus de Dirichlet conduisant à des distributions discrètes, la distribution des marqueurs a été décrite par un mélange de distributions normales dont les espérances étaient distribuées selon un processus de Dirichlet :

$$\begin{aligned} \text{nadir}_{jk}|\mathbf{y} &\hookrightarrow \mathcal{N}(\phi_{jk}, \sigma_j^2) \\ \phi_{jk}|G_{\phi_j} &\hookrightarrow G_{\phi_j} \\ G_{\phi_j} &\hookrightarrow \mathcal{DP}(M_j, G_{0j}) \end{aligned}$$

où nadir_{jk} est la valeur du nadir de PSA du k$^{\text{ième}}$ patient du groupe j (malade ou non malade), sachant ses mesures de PSA. Les mélanges de processus de Dirichlet correspondent à une approche semi-paramétrique. Une loi normale ($\mathcal{N}(m_j, \sigma^2_{\mu_j})$) a été considérée comme espérance a priori de la distribution des valeurs de ϕ_j (G_{0j}). Des a priori non informatifs ont été utilisés pour l'ensemble des paramètres du mélange de processus de Dirichlet :

$$\begin{aligned} m_j &\hookrightarrow \mathcal{N}(c_{0j}, 1/c_{1j}) \\ c_{0j} = 0 & \qquad c_{1j} = 0,001 \\ 1/\sigma^2_{\mu_j} &\hookrightarrow \text{Gamma}(\delta_{0j}/2, \delta_{1j}/2) \\ \delta_{0j} = 0,001 & \qquad \delta_{1j} = 0,001 \\ 1/\sigma_j^2 &\hookrightarrow \text{Gamma}(\varepsilon_{0j}/2, \varepsilon_{1j}/2) \\ \varepsilon_{0j} = 0,001 & \qquad \varepsilon_{1j} = 0,001 \\ M_j &\hookrightarrow \text{Gamma}(\alpha_{0j}, \alpha_{1j}) \\ \alpha_{0j} = 1 & \qquad \alpha_{1j} = 1 \end{aligned}$$

A une itération i de la chaîne MCMC, la probabilité qu'une valeur de nadir de PSA pour un nouveau patient du groupe j soit inférieure à un seuil donné c est donnée par :

$$\widetilde{P(\text{nadir}_j \leq c)} = \frac{\sum_{i=1}^{n_j} \Phi\left(\frac{c-\phi_{jki}}{\sigma_j^2}\right) + M_j \Phi\left(\frac{c-m_j}{\sigma^2_{\mu_j}}\right)}{M_j + n_j}$$

ϕ_{jki} correspond à la valeur de ϕ échantillonnée pour le k$^{\text{ième}}$ patient du groupe j à la i$^{\text{ème}}$ itération de la chaîne MCMC ; Φ dénote la fonction de répartition de la loi normale centrée réduite. Cette formule permet le calcul de la sensibilité et de la spécificité prédites à posteriori, donc la construction de la fonction d'utilité.

⊳ Simulations dans le cas de lois normales

Un ensemble de simulations a été réalisé afin de s'assurer de la validité de cette méthode. Elles reprennent le même scénario que celui décrit dans l'article accepté dans le Biometrical Journal, avec un marqueur dynamique suivant une loi normale chez les non malades ($\mathcal{N}(-0,3; 0,07^2)$) et chez les malades ($\mathcal{N}(-0,1; 0,07^2)$).

Deux cas de figure ont été analysés : un cas où le nombre total de patients était de 100 (50 patients par groupe) et un cas avec 200 patients (100 patients par groupe). A chaque fois, 5000 simulations ont été réalisées, permettant l'estimation du seuil optimal et de son intervalle de crédibilité par la méthode Bayésienne paramétrique, la méthode Bayésienne semi-paramétrique proposée précédemment et la méthode Bayésienne non paramétrique proposée dans l'article, appelée par la suite méthode non paramétrique empirique.

A l'issue de ces simulations, le biais relatif lié à l'estimation du seuil optimal par le mode, la moyenne et la médiane de la distribution a posteriori, ainsi que la probabilité de couverture et la largeur de l'intervalle de crédibilité à 95 % ont été calculés (tableau 5.5).

Tableau 5.5 – Comparaison des résultats des méthodes paramétrique, semi-paramétrique et non paramétrique empirique.

		Biais relatif			PC†		Largeur IC‡	
N	Méthode	Mode	Médiane	Moyenne	HDP	Quant	HDP	Quant
100	Paramétrique	-0,001	-0,002	-0,002	0,936	0,945	0,031	0,031
	SP⋆	-0,001	-0,001	-0,001	0,942	0,946	0,032	0,032
	NP§	-0,009	-0,011	-0,011	0,799	0,819	0,037	0,038
200	Paramétrique	-0,002	-0,002	-0,002	0,943	0,949	0,022	0,022
	SP⋆	-0,001	-0,001	-0,001	0,946	0,951	0,023	0,023
	NP§	-0,001	-0,003	-0,002	0,862	0,884	0,033	0,034

⋆ SP : semi-paramétrique ; §NP : non paramétrique ;

†PC : probabilité de couverture ; ‡IC : intervalle de crédibilité.

Ces résultats montrent que le biais relatif lié à l'estimation du seuil est faible (inférieur à 0,2 %) et similaire avec la méthode paramétrique et la méthode semi-paramétrique. De même, la probabilité de couverture est acceptable pour ces deux méthodes, même pour $N = 100$; les largeurs des intervalles de crédibilité sont proches. Ainsi, les résultats obtenus avec la méthode

semi-paramétrique sont aussi bons que ceux obtenus avec la méthode paramétrique lorsque les distributions du marqueur suivent des lois connues.

▷ Simulations dans le cas de mélanges de lois normales

Dans un second temps, la méthode semi-paramétrique a été évaluée lorsque la distribution du marqueur suit, dans l'un des deux groupes, un mélange de lois. Plus précisément, le cas où le marqueur suit un mélange de deux lois normales chez les malades a été considéré, la distribution chez les malades étant souvent un mélange de lois reflétant une hétérogénéité des patients par rapport au stade de la maladie. Pour simplifier les calculs, dans ces simulations, le marqueur ne correspondait pas à un marqueur calculé à partir des paramètres d'une trajectoire, mais à un marqueur directement mesuré.

Chez les malades, ce marqueur a été généré selon une loi normale $\mathcal{N}(-0,3\,;\,0,07)$; chez les non malades, il a été généré selon un mélange de lois normales : $0,5 \times \mathcal{N}(0,05;\sigma_{11}^2) + 0,5 \times \mathcal{N}(-0,25;\sigma_{12}^2)$. Trois paramètres ont été modifiés au cours des simulations :

- le nombre total de patients, N, prenant les valeurs 200 et 400, avec à chaque fois autant de patients dans les groupes malades et non malades ;
- les écart-types des deux lois normales constituant le mélange de distributions chez les malades : $\sigma_{11} = \{0,07\,;0,08\,;0,1\}$ et $\sigma_{12} = \{0,07\,;0,05\}$.

Pour chaque jeu de paramètres, 5000 simulations ont été réalisées, permettant le calcul du biais relatif lié aux différentes estimations ponctuelles du seuil, des probabilités de couverture et des largeurs des intervalles de crédibilité (tableau 5.6).

Le biais relatif des estimations ponctuelles diminue plus le nombre de patients augmente ; il est inférieur à 10 % pour 200 patients et inférieur à 8 % pour 400 patients, le mode de la distribution a posteriori conduisant en général à un biais un peu plus faible. La probabilité de couverture des intervalles de crédibilité tourne autour de 95 % ; les résultats obtenus sont similaires avec la méthode des quantiles ou la méthode HDP, de même en termes de largeur d'intervalles de crédibilité. La nouvelle méthode Bayésienne semi-paramétrique a donc globalement des propriétés tout à fait acceptables.

Ainsi, lorsque la distribution du marqueur ne correspond pas à une loi connue, il est possible d'utiliser la méthode Bayésienne non paramétrique empirique d'estimation du seuil pour des échantillons suffisamment grands. Pour des échantillons de taille modérée, la méthode Bayésienne semi-paramétrique fournit de bons résultats, même si elle est plus difficile à mettre en

Tableau 5.6 – Résultats de la méthode semi-paramétrique dans le cas de mélanges de lois chez les malades.

			Biais relatif			PC†		Largeur IC‡	
n	σ_{11}	σ_{12}	Mode	Médiane	Moyenne	HDP	Quant	HDP	Quant
400	0,07	0,07	0,003	0,034	0,021	0,949	0,952	0,065	0,068
400	0,08	0,05	0,036	0,083	0,053	0,945	0,931	0,096	0,105
400	0,10	0,05	0,028	0,064	0,042	0,947	0,928	0,080	0,086
200	0,07	0,07	0,011	0,051	0,034	0,938	0,950	0,087	0,091
200	0,08	0,05	0,055	0,105	0,075	0,930	0,924	0,113	0,120
200	0,10	0,05	0,043	0,084	0,061	0,929	0,917	0,096	0,102

†PC : probabilité de couverture ; ‡IC : intervalle de crédibilité.

œuvre ; cette méthode constitue une solution plus prudente vis-à-vis de la méthode paramétrique, car il y a souvent une incertitude sur le choix de la distribution du marqueur.

Cette dernière approche permet d'obtenir une modélisation très souple des distributions des marqueurs ; pour autant, il se peut que la distribution obtenue reflète des caractéristiques propres uniquement à l'échantillon considéré et non à la population globale pour laquelle le marqueur est destiné. Il faut donc trouver un équilibre entre souplesse et capacité à être généralisable à la totalité de la population dont est issu l'échantillon.

▷ Application au nadir de PSA

La méthode semi-paramétrique a été appliquée à l'estimation du seuil optimal du nadir de PSA et à son intervalle de crédibilité. Les résultats obtenus pour $r = 0,5$ sont présentés dans le tableau 5.7, ainsi que ceux qui avaient été obtenus en utilisant des distributions log normales ou des valeurs extrêmes.

Le mode, la médiane et la moyenne de la distribution a posteriori du seuil optimal obtenus avec la méthode semi-paramétrique se situent entre ceux obtenus par la méthode paramétrique avec des lois log normales et des valeurs extrêmes. Les intervalles de crédibilité sont plus larges qu'avec la méthode paramétrique, ce qui est assez cohérent.

Globalement, quelle que soit la méthode retenue, les résultats obtenus en termes d'estimation ponctuelle sont assez similaires.

Tableau 5.7 – Comparaison des estimations de seuil optimal de nadir de PSA obtenues pour $r = 0,5$ avec la loi log normale, la loi des valeurs extrêmes et la méthode semi-paramétrique.

Méthode	Mode	Médiane	Moyenne	IC* Quant,	IC* HDP
Log normale	0,273	0,291	0,296	[0,221 ; 0,392]	[0,219 ; 0,388]
Valeurs extrêmes	0,230	0,241	0,245	[0,173 ; 0,337]	[0,166 ; 0,324]
Semi-paramétrique	0,250	0,256	0,264	[0,159 ; 0,423]	[0,142 ; 0,386]

* IC : intervalle de crédibilité

5.6 Bilan du chapitre 5

Dans cette partie, un ensemble de méthodes Bayésiennes d'estimation du seuil optimal d'un marqueur a été développé, permettant d'obtenir une estimation ponctuelle et un intervalle de crédibilité de ce seuil. Dans le cas de la méthode Bayésienne paramétrique, il suffit de connaître la loi du marqueur dans les deux groupes et de savoir échantillonner dans la distribution a posteriori des paramètres de la distribution du marqueur pour obtenir le seuil optimal. Ces contraintes sont totalement supprimées dans le cas de la méthode Bayésienne semi-paramétrique, applicable quelle que soit la distribution du marqueur, au prix de calculs plus complexes. Par rapport aux méthodes existantes, la nouveauté est de pouvoir fournir un intervalle de crédibilité correct dans toutes les situations, même les plus complexes.

La méthode Bayésienne d'estimation du seuil permet de tenir compte, au niveau de l'estimation ponctuelle, de l'incertitude quant aux valeurs des paramètres de la distribution du marqueur. Ceci n'est pas le cas de la plupart des méthodes existant actuellement ; lorsque la taille de l'échantillon est grande, les résultats sont quasiment identiques, mais ce n'est plus vrai pour des effectifs réduits. De plus, dans le cas d'un marqueur dynamique estimé à partir de données longitudinales, la méthode Bayésienne permet facilement de tenir compte de l'incertitude quant à l'estimation des valeurs de marqueurs dans la détermination du seuil optimal et de son intervalle de crédibilité.

Il est tout de même à noter que l'intervalle de crédibilité fourni correspond à celui du seuil optimal ; en s'éloignant un peu de cette valeur, l'utilité obtenue est plus faible, mais peut, dans certains cas, rester proche de l'utilité du seuil optimal ; dans d'autres cas, elle peut à l'inverse diminuer très rapidement. L'intervalle de crédibilité fourni ne reflète pas les valeurs de seuils qui peuvent être proche du seuil optimal en termes d'utilité.

Dans toute cette partie, il n'a été tenu compte que de l'incertitude sur les valeurs de paramètres de la distribution du marqueur, en négligeant l'incertitude sur l'estimation de la prévalence. Cette incertitude aurait pu facilement être introduite en échantillonnant des valeurs dans la distribution a posteriori de la prévalence, puis en intégrant ces valeurs à la chaîne MCMC permettant le calcul des fonctions d'utilité. Etant donné qu'il n'existe pas vraiment d'information a priori sur la prévalence dans la littérature, un a priori non informatif sous la forme d'une loi gamma (Gamma$(0,001; 0,001)$) aurait pu être choisi, conduisant dans ce cas à un écart type de la distribution a posteriori de la prévalence de 0,03, relativement faible, et donc négligeable dans l'estimation du seuil optimal. L'intégration de l'incertitude sur la valeur de la prévalence aurait pu être intéressante si une information a priori sur celle-ci avait été disponible.

Toute la difficulté de ce chapitre a été de développer une méthode générale pour l'estimation de l'intervalle de crédibilité du seuil optimal d'un marqueur. Ceci peut être vu comme une lubie des statisticiens, mais l'intervalle de crédibilité a des conséquences sur la façon d'utiliser le seuil optimal du marqueur. Par exemple, pour $BN/CN = 2$, l'estimation ponctuelle du seuil de nadir était de 0,135 ng/mL d'après la moyenne de la distribution a posteriori, l'intervalle de crédibilité à 95 % étant de $[0,077; 0,185]$ avec la méthode des quantiles. Pour un patient dont le nadir est de 0,05 ng/mL et pour lequel la valeur $BN/CN = 2$ semble raisonnable, alors le médecin peut en toute confiance ne pas proposer de biopsie pour le patient. A l'inverse, si l'intervalle de crédibilité était de $[0,010; 0,300]$, la réalisation d'une biopsie ne serait peut-être pas déconseillée. Le fait de fournir un intervalle de crédibilité aux estimations de seuil permet de justifier l'intérêt de leur définition. Ainsi, un intervalle de crédibilité de $[0,001; 5,000]$ indiquerait, dans le cadre du nadir de PSA, que l'utilisation du seuil obtenu en tant que tel n'a pas beaucoup d'intérêt, vu l'incertitude quant à son estimation. Les intervalles de crédibilité de nadir de PSA sont dans l'ensemble assez resserrés par rapport à l'étendue des valeurs possibles de nadirs, justifiant ainsi la définition de seuils de nadirs.

De manière plus générale, la méthode Bayésienne d'estimation du seuil pourrait être généralisée à tous les problèmes d'optimisation en présence d'incertitude sur certains paramètres de la fonction a optimiser. En échantillonnant les valeurs de paramètres dans leur distribution a posteriori, il est possible d'obtenir la distribution a posteriori de la fonction à optimiser et celle du paramètre d'intérêt. Par exemple, la plupart des méthodes Bayésiennes développées pour optimiser un plan d'expérience reposent sur le calcul de la fonction d'utilité moyenne et sur la détermination de la valeur qui optimise cette fonction moyenne (Chaloner et Verdinelli, 1995). Les distributions a posteriori des paramètres sont souvent approximées par des lois normales

lorsque le calcul de l'intégrale pour la fonction moyenne n'est pas réalisable de façon analytique. Muller et Parmigiani (1996) ont également proposé, dans ce cas, de réaliser l'intégration par la méthode de Monte Carlo, ce qui revient à la méthode proposée par Wang et Geisser (2005) pour l'estimation du seuil optimal lorsque les distributions a priori des paramètres ne sont pas conjuguées. La méthode Bayésienne proposée précédemment dans le cadre du seuil d'un marqueur pourrait permettre d'obtenir facilement une estimation du critère d'intérêt et de son intervalle de crédibilité. Cette méthode est donc généralisable à de très nombreux problèmes.

Enfin, la détermination de plusieurs seuils de nadir de PSA permet d'individualiser la prise de décision en fonction des préférences du patient en termes de qualité ou de quantité de vie, ce qui constitue un réel apport.

Chapitre

6 Variabilité des préférences et choix du seuil

La détermination du seuil optimal d'un marqueur dépend de la sensibilité, de la spécificité, de la prévalence et du ratio bénéfice net sur coût net (ou risque de maladie à partir duquel le patient accepte d'être traité). Ce dernier terme dépend des préférences des patients et des cliniciens.

Dans le cas de la détection de la persistance du cancer de la prostate, même pour un risque de maladie de 10 %, un patient très craintif vis-à-vis de la maladie va peut être opter pour une biopsie. Pour ces patients, un seuil bas de nadir de PSA doit être retenu afin de favoriser la sensibilité du test. A l'inverse, une personne craignant un peu moins la maladie mais beaucoup plus la réalisation d'une biopsie ne va accepter celle-ci qu'à partir d'un risque de maladie de 30 à 40 % par exemple. Dans ce cas, la spécificité est à privilégier, en retenant un seuil relativement élevé de nadir de PSA (ces valeurs de risque ne sont utilisées qu'à titre d'exemple, et ne correspondent pas forcément à celles des patients).

De la même façon, certaines équipes de soins souhaitent pouvoir retraiter le plus tôt possible les patients pour lesquels des cellules cancéreuses persistantes sont détectées. En effet, le second traitement UFHI est d'autant plus efficace que le taux de PSA au moment du traitement est faible, nécessitant donc que le second traitement soit réalisé rapidement après le premier (Murat et *al.*, 2010). Ces équipes souhaitent donc privilégier la sensibilité du nadir de PSA, quitte à réaliser des biopsies à tort. A l'inverse, d'autres équipes préfèrent limiter le nombre de biopsies ; dans ce cas, la spécificité du nadir de PSA est plus importante. Ainsi, il existe vraisemblablement dans la population des variations au niveau des préférences quant à la réalisation ou non de

biopsies, ce qui se traduit par des variations au niveau du risque de maladie à partir duquel un clinicien propose ou un patient accepte une biopsie.

Deux approches ont été envisagées afin de refléter ces variations dans le choix du seuil optimal du nadir de PSA. La première consiste à éliciter les souhaits d'un grand nombre de personnes en termes de risque de maladie à partir duquel une biopsie est acceptée, afin de construire la distribution de probabilité de cette valeur. Cette distribution peut être introduite dans la chaîne MCMC de l'estimation du seuil comme une source supplémentaire d'incertitude, au même titre que l'incertitude concernant les valeurs de paramètres de distribution du marqueur. La seconde approche consiste à proposer un ensemble de seuils en fonction des préférences des cliniciens et des patients, correspondant ainsi à une analyse de sensibilité.

6.1 Elicitation de la distribution de probabilité du risque de maladie pour la biopsie

Dans cette première partie, l'objectif est de déterminer la distribution de probabilité du risque de maladie à partir duquel une biopsie est recommandée ou acceptée – appelé par la suite risque de maladie pour la biopsie – et d'intégrer cette source de variabilité dans la détermination du seuil optimal. Des méthodes ont été proposées, au cours de la partie 2.2.4, pour estimer les différentes utilités intervenant dans le calcul du risque de maladie justifiant la biopsie pour un patient ou un ensemble de patients. Néanmoins, la formalisation des préférences d'un patient en termes de qualité ou de quantité de vie reste une opération difficile. Dans ce type de situation, au lieu de retenir une unique valeur de préférence, il peut être plus judicieux de décrire l'incertitude du patient, ou d'un ensemble de patients, concernant ces préférences. Ce type d'opération fait appel à l'élicitation d'information.

6.1.1 Principe de l'élicitation

L'objet de l'élicitation est de décrire l'incertitude concernant la valeur d'une quantité inconnue à partir d'une distribution de probabilité représentant correctement la connaissance – et l'incertitude – d'un expert ou d'un ensemble d'experts sur cette question. Elle est utilisée dans deux domaines principaux. Le premier concerne les projets d'ingénierie complexes, quasiment uniques, où il est difficile de déterminer les performances combinées ou individuelles des différents composants d'un système, comme par exemple dans les installations nucléaires. L'élicitation joue

également un rôle important dans le domaine de la prise de décision, lorsque les conséquences d'une action sont difficiles à quantifier.

Prenons l'exemple d'un marqueur diagnostic dont le résultat est directement utilisé pour décider de la réalisation d'un traitement, comme une opération chirurgicale. Un clinicien peut être dubitatif quant au risque de maladie à partir duquel il va conseiller l'opération, cette incertitude provenant de deux sources différentes. La première est l'incertitude *aléatoire*, liée au fait que le bénéfice d'un traitement peut varier d'un patient à l'autre. La seconde source d'incertitude, dite *épistémique*, reflète le fait que le clinicien ne connaît pas de façon exacte le bénéfice attendu suite à la réalisation de l'opération.

Ces deux sources d'incertitude peuvent être distinguées à partir d'un exemple statistique. Lorsqu'une régression linéaire est réalisée, le modèle aléatoire s'écrit :

$$y = \alpha + \beta x + \sigma\varepsilon$$

Les résidus (ε) représentent l'incertitude aléatoire ; les paramètres du modèle (α, β et σ) sont liés quant à eux à l'incertitude épistémique. Ces deux sources d'incertitude sont intégrées sans distinction dans la distribution de probabilité issue de l'élicitation.

Pour l'exemple du risque de maladie à partir duquel une opération est conseillée, un clinicien peut, lors d'une discussion informelle, indiquer que ce risque se situe entre 20 et 40 %, mais qu'il est certainement plus proche de 40 %. Toute la difficulté de l'élicitation est de retranscrire cette information en une distribution de probabilité.

L'objectif de cette partie n'est pas d'effectuer une revue complète de la littérature concernant l'élicitation d'information, mais de décrire les principales méthodes existantes, en insistant particulièrement sur celles utiles dans le cas de l'élicitation de l'information concernant un risque de maladie pour prendre une décision. L'article de Garthwaite et *al.* (2005) et l'ouvrage de d'O'Hagan et *al.* (2006) constituent une bonne introduction à ce sujet.

6.1.2 Elicitation d'information à partir de l'avis d'un expert

Deux grands types d'approches ont été décrits pour éliciter l'information provenant d'un expert : l'une est non paramétrique ou semi-paramétrique, l'autre paramétrique. La première méthode fait l'hypothèse que la fonction de répartition de la quantité inconnue est continue et non irrégulière, de sorte qu'elle peut être facilement reproduite à partir d'un nombre fini de quantiles. Très souvent, les quartiles de la distribution sont demandés, car il est difficile d'obtenir

une information plus précise. Une autre méthode peut consister à demander les deux valeurs extrêmes de la quantité, à diviser l'intervalle ainsi obtenu en sous intervalles, puis à demander la probabilité que la valeur soit contenue dans chacun des sous intervalles. Cette méthode nécessite de poser de nombreuses questions à l'expert pour représenter correctement la distribution de probabilité. De plus, la distribution ainsi obtenue est difficilement exploitable dans une chaîne MCMC.

L'approche paramétrique d'élicitation d'information repose sur l'hypothèse que l'opinion peut être représentée par un membre d'une famille de distributions paramétriques. L'élicitateur a besoin de poser moins de questions à l'expert pour déterminer les paramètres de la distribution. Une fois la distribution construite, quelques questions supplémentaires peuvent permettre de valider l'opération d'élicitation en comparant les valeurs prédites par la distribution et les valeurs données par l'expert. Le résultat ainsi obtenu est directement exploitable dans une chaîne MCMC. Les distributions les plus couramment utilisées sont les distributions triangulaires, normales ou beta. L'élicitation de quelques quantiles de la distribution permet l'estimation des paramètres de ces lois.

6.1.3 Combinaison de l'information issue de plusieurs experts

Il est souvent souhaitable – et recommandé – de demander l'avis de plusieurs experts, médecins ou patients, pour construire la distribution de probabilité de la quantité d'intérêt. Une difficulté consiste alors à concilier leurs avis. Deux grandes familles de méthodes ont été proposées : l'agrégation mathématique de l'information, ou l'agrégation dite des " comportements ". La seconde méthode consiste à rassembler les différents experts et à aboutir, grâce à un débat, à un consensus sur la distribution de probabilité. Ce type de stratégie est présenté par Clemen et Winkler (1999) et O'Hagan et *al.* (2006) et n'est pas détaillé par la suite. L'agrégation mathématique consiste à recueillir séparément les opinions des experts, puis à les combiner a posteriori.

6.1.3.1 L'agrégation mathématique automatique

Une première façon de combiner les distributions obtenues auprès des différents experts est d'en effectuer la moyenne arithmétique. La distribution obtenue représente ainsi l'avis de l'ensemble des experts. Il est également possible d'effectuer la moyenne logarithmique des différentes distributions. Dans ce cas, une valeur de la quantité étudiée qui n'est soutenue que par un seul expert est rejetée dans le résultat final ; c'est le principe de l'exclusion. La distribution

ainsi construite est souvent unimodale et moins étendue que celle obtenue avec la moyenne arithmétique.

Des variantes de ces deux méthodes ont été proposées, notamment en affectant des poids à chaque expert, poids représentant le degré de confiance en l'avis de chacun d'eux, mais ces poids sont difficilement quantifiables (O'Hagan et *al.*, 2006). L'inconvénient majeur de ces méthodes est qu'elles ne tiennent pas compte de la dépendance entre les informations fournies par les experts. Par exemple, deux cliniciens peuvent avoir reçu la même formation ; ils auront vraisemblablement des avis similaires. En considérant les avis des experts comme indépendants, les méthodes d'agrégation automatique surestiment la quantité d'information dans la distribution finale (Clemen et Winkler, 1999 ; Garthwaite et *al.*, 2005).

6.1.3.2 La méthode du " supra Bayesian "

L'approche dite du "supra Bayesian" fait appel aux principes de l'inférence Bayésienne ; elle a été introduite par Morris (1974 ; 1977). Le décideur propose une distribution de probabilité de base $P(\theta)$ pour la quantité inconnue θ, puis recueille l'avis des différents experts sur la question sous la forme de distributions de probabilité $\mathbf{f}$. Il peut alors améliorer sa connaissance sur la quantité inconnue à l'aide du théorème de Bayes :

$$P(\theta|\mathbf{f}) \propto P(\mathbf{f}|\theta) \times P(\theta)$$

$P(\mathbf{f}|\theta)$, la fonction de vraisemblance, correspond à la distribution de probabilité conjointe des opinions des experts, tenant compte de la précision et du biais des différents avis, ainsi que de leurs dépendances. Elle est souvent difficile à spécifier.

Une première approche consiste à supposer que l'avis des n experts peut être transcrit sous la forme de lois normales de moyenne μ_i et de variance σ_i^2 (Winkler, 1981). Le vecteur des moyennes $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\}$ représente l'estimation de θ selon les experts. L'erreur commise est décrite par le vecteur d'erreurs : $\boldsymbol{\varepsilon} = \{\mu_1 - \theta, \ldots, \mu_n - \theta\}$, supposé suivre une distribution normale multivariée de moyenne le vecteur $(0, \ldots, 0)^T$ (si les avis des experts sont non biaisés) et de matrice de variance covariance $\boldsymbol{\Sigma}$. Cette matrice est soit déterminée à partir des données (Winkler, 1981), soit fixée de manière subjective. En utilisant des a priori non informatifs, la distribution a posteriori de θ correspond à une loi normale. La prise en compte de la dépendance entre les avis des experts reste tout de même difficile ; de plus, les distributions obtenues sont forcément normales et unimodales.

Une approche plus générale pour modéliser la dépendance entre les avis a été proposée par Jouini et Clemen (1996). Elle repose sur l'utilisation de " copulas ". La fonction de vraisemblance décrivant les dépendances entre les opinions est exprimée grâce à une fonction copula qui combine les distributions marginales correspondant aux avis des experts. Par exemple, si l'information obtenue auprès de l'expert i correspond à une distribution de probabilité de densité f_i et de fonction de répartition F_i et si le décideur a un a priori non informatif concernant θ, alors la distribution a posteriori de θ est donnée par :

$$P(\theta|f_1,\ldots,f_n) \propto C[1-F_1(\theta),\ldots,1-F_n(\theta)]\prod_{i=1}^{n} f_i(\theta)$$

où C correspond à une fonction copula. En utilisant la famille de fonctions copula de Frank, cette fonction est paramétrable à partir du taux de concordance entre les experts. Ce taux de concordance peut être obtenu en posant aux experts un ensemble de questions sur un domaine similaire, et en mesurant la concordance dans les réponses. Cette méthode permet de modéliser des types de relations de dépendance complexes ; les distributions a posteriori peuvent ainsi être très variées et non forcément unimodales.

Les deux dernières approches sont particulièrement utiles pour combiner les avis de différents experts (médecins et patients étant ici considérés comme les experts) ; elles permettent d'obtenir une distribution qui est directement exploitable pour l'estimation du seuil optimal du nadir de PSA.

6.1.4 Intégration de l'information élicitée dans la détermination du seuil optimal

Dans la formule (5.8), le ratio bénéfice net sur coût net (relié au risque de maladie pour la biopsie selon la relation $BN/CN = (1-r)/r$) peut être considéré comme une variable aléatoire dont la distribution de probabilité est obtenue à partir des résultats de la phase d'élicitation. A chaque itération de la chaîne MCMC concernant la distribution des marqueurs, une valeur de risque de maladie pour la biopsie peut être échantillonnée dans sa distribution de probabilité, puis intégrée à la chaîne MCMC pour le calcul de l'utilité espérée.

Le seuil optimal obtenu et son intervalle de crédibilité tiennent compte à la fois des incertitudes statistiques liées à l'estimation des paramètres de distribution du marqueur (et éventuellement de la prévalence), ainsi que de la variabilité des préférences des cliniciens en

termes de sensibilité et de spécificité et des patients en termes de qualité ou de quantité de vie, et ce, bien que ces deux sources de variabilité soient de natures totalement différentes.

Le seuil obtenu est un seuil moyen sur l'ensemble des préférences. Il est utile si ces variations de préférences sont faibles d'un clinicien ou d'un patient à l'autre, c'est à dire si elles sont essentiellement liées au fait que les personnes ne donnent pas la même valeur quantitative au même niveau de risque. Dans certains cas, les préférences sont très variables d'un patient à l'autre ou d'une équipe de soins à l'autre, car les volontés sont très différentes. La définition d'un seuil unique a moins de sens, car il n'est plus possible de définir pour qui ce seuil est réellement utile. Dans ce type de situation, il semble plus raisonnable d'établir un ensemble de seuils selon les niveaux de risque de maladie à partir desquels une biopsie sera conseillée.

Dans le cadre des données de PSA après traitement UFHI, il semble que les souhaits des équipes de soins en termes de sensibilité et de spécificité soient très variables. De même, les craintes vis-à-vis de la maladie et de la réalisation d'une biopsie varient de manière non négligeable d'un patient à l'autre. Ainsi, il n'était pas adapté d'intégrer la distribution du risque de maladie pour la réalisation de biopsie dans l'estimation du seuil. C'est pourquoi la seconde stratégie, consistant à définir différents seuils de nadir de PSA, a été privilégiée.

6.2 Analyse de sensibilité en fonction des préférences individuelles

6.2.1 Une utilité collective pour des décisions individuelles

Dans cette approche, plusieurs valeurs de seuil optimal de nadir de PSA ont été définies en fonction des valeurs de risque de maladie à partir desquelles un clinicien proposera une biopsie et un patient sera susceptible de l'accepter. Pour un patient, ce niveau de risque est influencé par des préférences en termes d'état de santé, ainsi que par son aversion pour le cancer de la prostate, d'une part, et la réalisation de biopsies, d'autre part.

Le fait de définir plusieurs seuils de PSA en fonction du niveau de risque de maladie pour la réalisation de biopsies permet d'individualiser la prise de décision. Pour une personne très âgée, le risque de maladie retenu peut être surélevé, afin de tenir compte de la pénibilité d'une biopsie et du peu de bénéfice attendu d'un éventuel second traitement à un âge élevé. De même, le niveau de risque peut dépendre de l'état de santé général. Ainsi, en fonction de certaines

caractéristiques du patient et de ses préférences, le niveau de risque, et donc le seuil de nadir de PSA, peut être adapté.

Le seuil optimal est le seuil maximisant l'utilité espérée lorsque ce seuil est utilisé dans une population ; c'est un seuil de décision, mais sa valeur de positivité peut être définie à l'échelle du patient, en tenant compte du risque individuel de maladie à partir duquel une biopsie serait acceptée.

Néanmoins, tous les seuils ne sont pas forcément utiles. Cette notion est abordée dans l'article rédigé pour The European Urology, dont les méthodes principales sont présentées ci-après.

6.2.2 Méthodes

6.2.2.1 Limites hautes et basses de risque de maladie pour la réalisation d'une biopsie

Pour un risque très faible de maladie pour la réalisation de biopsie, correspondant à des patients très craintifs vis-à-vis du cancer, l'utilité attendue liée à la stratégie " effectuer une biopsie en fonction du nadir de PSA " risque de ne pas être meilleure que celle consistant à effectuer une biopsie systématiquement, indépendamment des valeurs de PSA. En effet, le coût d'effectuer une biopsie à tort est jugé dans ce cas négligeable par rapport au bénéfice d'effectuer une biopsie à raison. A l'inverse, pour des risques très élevés de maladie à partir desquels la biopsie est acceptée, correspondant aux patients très craintifs vis-à-vis des biopsies, l'utilité attendue de la stratégie " effectuer une biopsie en fonction du nadir de PSA " risque de ne pas être meilleure que la stratégie " ne jamais effectuer de biopsie ", le coût d'effectuer une biopsie à tort étant dans ce cas jugé très élevé.

Ainsi, il est nécessaire de définir des valeurs limites de risque de maladie pour effectuer la biopsie : une limite basse en dessous de laquelle il vaut mieux effectuer systématiquement une biopsie et une limite haute au dessus de laquelle il est déconseillé d'effectuer une biopsie, quelles que soient les valeurs de nadir de PSA. Les seuils de nadir de PSA ne sont alors calculés que pour des risques compris entre ces bornes. La détermination des limites a été effectuée en comparant le bénéfice espéré des trois stratégies mentionnées ci-dessus à l'aide de courbes de décision, qui avaient été présentées précédemment pour le choix de marqueurs (chapitre 3).

Vickers et Elkin (2006) comparent les bénéfices espérées de la réalisation d'une biopsie. Pour la stratégie " effectuer une biopsie suivant la valeur du nadir de PSA ", ce bénéfice est

donné dans le cas d'un risque de maladie r pour la réalisation d'une biopsie par :

$$\text{bénéfice espéré}(r) = \text{Sen}(\hat{c}_r)\pi - (1 - \text{Spe}(\hat{c}_r))(1 - \pi)\frac{r}{1 - r}$$

$\hat{c}_r$ correspond au seuil optimal de nadir de PSA estimé pour la valeur r. Pour la stratégie " ne jamais effectuer de biopsie ", le bénéfice espéré est nul, puisque qu'aucune biopsie n'est réalisée. Enfin, pour la stratégie " effectuer systématiquement une biopsie ", le bénéfice espéré d'une biopsie pour une valeur r est :

$$\text{bénéfice espéré}(r) = \pi - (1 - \pi)\frac{r}{1 - r}$$

la sensibilité de cette stratégie étant de 1 et la spécificité étant nulle. Les intersections entre ces courbes permettent de déterminer les limites hautes et basses de risque de maladie pour la réalisation d'une biopsie pour lesquelles l'utilisation des valeurs de nadir de PSA est bénéfique.

En réalité, il existe une incertitude sur la valeur du seuil de nadir qui optimise l'utilité à un niveau de risque donné. En conséquence, les intervalles de crédibilité à 95 % du seuil optimal pour chaque valeur de r ont été utilisés pour tracer les intervalles de crédibilité à 95 % du bénéfice espéré de la biopsie en utilisant la stratégie reposant sur les valeurs de PSA. La limite basse de risque a été définie comme étant la moyenne des valeurs de risque pour lesquelles l'intervalle de crédibilité du bénéfice espéré lié à la stratégie basée sur les PSA englobait la courbe liée à la stratégie biopsie systématique. Pour la limite haute, la même technique a été utilisée, mais en considérant la stratégie " aucune biopsie ".

6.2.2.2 Correspondance par rapport au nadir clinique

Tout le travail réalisé sur l'estimation du seuil optimal du nadir de PSA a été fondé sur les valeurs de nadirs de PSA issues de la modélisation robuste des mesures longitudinales, afin de s'affranchir des problèmes de fréquence de mesures et de valeurs aberrantes. Néanmoins, en pratique courante, le nadir n'est pas obtenu à partir de la modélisation des profils de PSA, mais à partir de la plus faible mesure de PSA observée. De plus, la modélisation repose sur l'ensemble du suivi de chaque patient en s'arrêtant à la date de première biopsie positive ou de dernière biopsie. Dans la réalité, les cliniciens ne souhaitent pas attendre la première biopsie positive pour calculer le nadir, car il est alors trop tard. Typiquement, le nadir clinique est défini comme étant la plus basse valeur de PSA mesurée au cours des trois premiers mois de suivi.

Il se peut que les performances diagnostiques des seuils de nadir estimés à partir des nadirs modélisés soient différentes de celles obtenues en retenant ces mêmes seuils de positivité, mais en définissant le statut du patient à partir du nadir clinique. Une comparaison des performances " théoriques " et des performances " cliniques " a donc été réalisée dans l'article rédigé pour The European Urology.

6.2.2.3 Redéfinition du statut des patients

Pour 40 % des patients déclarés comme malades, la première biopsie positive a été enregistrée au delà d'un an de suivi. La population de patients considérés comme " malades " était donc constituée d'un mélange :

- de patients pour lesquels il restait des cellules cancéreuses actives non traitées par le traitement UFHI ;
- et de patients pour lesquels le traitement avait été efficace, mais qui avaient effectué une récidive.

Les nadirs de PSA des patients du premier groupe étaient plus élevés que ceux du second groupe. En réalité, l'utilisation du nadir de PSA pour déclencher les biopsies vise plus à rechercher les patients pour lesquels la persistance de cellules cancéreuses est détectable précocement, afin d'effectuer un second traitement UFHI le plus tôt possible. Pour les patients pour lesquels la biopsie est positive tardivement, d'autres tests diagnostiques, plus tardifs, ont été proposés par Blana et *al.* (2009).

Ainsi, dans l'article rédigé pour The European Urology, le statut des patients a été redéfini, en marquant comme malades les patients ayant une biopsie positive durant la première année de suivi, les patients restants étant considérés comme des non malades (biopsie positive après un an ou sans biopsie positive). Ceci a conduit à une prévalence de 31 % (90 patients malades sur 289), plus faible que celle obtenue avec la première définition du statut des patients. A chaque itération de l'algorithme MCMC, les nadirs de PSA chez les non malades suivaient toujours des distributions log normales ; chez les malades, les nadirs suivaient également une distribution log normale pour les valeurs de logarithme de nadir en dessous de 1. Ainsi, le seuil optimal de nadir de PSA a pu être estimé à partir des mêmes méthodes que celles détaillées dans le chapitre 5.

6.2.3 Article rédigé pour The European Urology

L'article rédigé pour The European Urology définit une stratégie de déclenchement de biopsies après traitement UFHI reposant sur les valeurs de nadirs de PSA. Cette stratégie est

valable pour les patients ayant plus de 90 jours de suivi, et pour lesquels il existe déjà une suspicion de persistance de cellules cancéreuses. Les seuils de nadir de PSA ont été définis en fonction du risque de persistance de cellules cancéreuses au dessus duquel un clinicien conseillerait la réalisation d'une biopsie à un patient, patient répondant aux critères évoqués ci-dessus.

6.2.3.1 Article

PSA nadir to initiate control biopsies after a first High-Intensity Focused Ultrasound (HIFU) session for localised persistent prostate cancer

Fabien Subtil, Sébastien Crouzet, François-Joseph Murat, Albert Gelet, Muriel Rabilloud

Corresponding author:

Fabien Subtil

Hospices Civils de Lyon - Service de Biostatistique

162 avenue Lacassagne

F-69003 Lyon France

Phone: (+33) 4 72 11 57 51

Fax: (+33) 4 72 11 51 41

E mail: fabien.subtil@chu-lyon.fr

Keywords:

Diagnostic test; High-intensity focused ultrasound; Prostate cancer; Prostate specific antigen nadir; Salvage therapy; Treatment outcome.

Word count of text: 2321 words.

Word count of abstract: 249 words.

Summary

Background: One of the main advantages of the high-intensity focused ultrasound (HIFU) therapy for localised prostate cancer (PCa) is that it can be repeated, but the criteria to trigger biopsies to check for persistent PCa are still debated.

Objectives: Define the prostate specific antigen (PSA) nadir value above which biopsies should be performed after a first HIFU session to early diagnose local persistent PCa –not recurrent PCa– considering the risk not to diagnose it vs. the risk of performing biopsy.

Design, Setting, and Participants: Retrospective study of 289 patients seen between 2000 and 2007 who underwent HIFU for PCa and who had at least one biopsy past 90 d after treatment

Intervention: None

Measurements: The nadir was calculated for each patient using a mixed model describing individual longitudinal PSA measurements. Several PSA nadir thresholds were defined, depending on the risk of local persistent PCa above which a physician is willing to perform a biopsy.

Results and limitations: Initiating biopsy only in patients whose PSA nadir is above the defined nadir thresholds would lead to 36 to 58% proportion of positive biopsies, 11 to 30% rate of missed cases, and 4 to 81% of patients with biopsy. The results are dependent on the definition of local persistent PCa.

Conclusion: Using the PSA nadir criterion to initiate biopsies leads to earlier salvage treatments for local persistent PCa after HIFU, with a substantial reduction of the number of biopsies performed compared to systematic biopsy strategy, and low rates of missed cases.

Take home message (35 words to be deleted from the final manuscript)

After HIFU for prostate cancer, the triggering of biopsy on PSA nadir values allows earlier salvage treatments, important reductions of the number of biopsies made compared to the systematic biopsy strategy, and few missed cases.

1. Introduction

High-intensity focused ultrasound (HIFU) is an efficient treatment for patients with low- or intermediate-risk localised prostate cancer (PCa) with a low associated morbidity [1, 2]. It is now considered as an option by the French Association of Urology for stage T1-2 disease, in men age ≥70 y, with a PSA level <15 ng/ml and a Gleason score <7 [3]. Five-year disease-free survival rates of 66 to 77% have been reported [4-6]. One interest of HIFU treatment is that it can be repeated [6] in case of local persistent PCa, though other salvage treatments, such as external beam radiotherapy (EBRT) [7] can be also considered. However, these salvage treatments make it necessary to diagnose early local persistent PCa. Indeed, though residual cancer can be detected using sextant biopsies, possibly guided by dynamic contrast-enhanced MRI [8], there is still a clear need for non-invasive diagnostic tools to initiate these biopsies.

A PSA nadir below 0.2 ng/ml was found to be a good prognostic factor [9, 10]. With the aim of taking a decision about the realisation of a biopsy for a specific patient, Blana et *al.* [5] proposed diagnostic rules based on PSA threshold values, 'PSA nadir plus', PSA velocity, and PSA doubling times; they suggested that the biochemical failure should be defined by an increase in PSA level of 1.2 ng/ml over the nadir. In their analysis, Blana et *al*. made no distinction between local persistent and local recurrent PCa. Earlier-applicable diagnostic rules were suggested, that were more oriented toward the detection of local persistent PCa. One of the studies compared the diagnostic accuracy of the PSA nadir vs. the time to that nadir and found that the former was a better diagnostic criterion [11]. However, to be used as a diagnostic test, a PSA nadir threshold above which a biopsy should be performed has to be defined taking into account the patient aversion toward prostate cancer and toward biopsy, and the physician's overall preferences in terms of expected proportions of positive biopsies.

This is the focus of the present article meant to help clinicians in their decision to propose a biopsy for the diagnosis of local persistent PCa after HIFU.

2. Materials and methods

2.1. Patients

The study involved all patients treated at Edouard Herriot hospital (Lyon, France) between 2000 and 2007 with the second-generation HIFU device (EDAP TMS S.A., Vaulx-en-Velin, France). Among them, 289 patients were retrospectively included on the following criteria: no previous hormonal therapy or other treatment for PCa, at least five PSA measurements with one after 90 d, and at least one biopsy after 90 d.

2.2. Follow-up

In all patients, PSA measurements were performed every three weeks during the first four mo, then monthly until the eighth mo, then every four mo. At least one prostate biopsy was taken as control between three and six mo after HIFU and additional ones in case of rising PSA value. Patients were considered as having a local persistent cancer after HIFU treatment if they had at least one positive biopsy between 90 d and one yr. PSA measurements taken after the last biopsy in patients without local persistent cancer or after the first positive biopsy in the others were discarded.

2.3. Statistical analysis

PSA longitudinal measurements were modelled using a non-linear mixed model described by Subtil and Rabilloud [11]. To reduce the impact of the high biological variability of PSA values, the PSA nadir and the time to that nadir were calculated for each patient from that

model [12]. In some cases, there was no nadir; the last PSA measurement was then considered as a nadir estimate.

The optimal PSA nadir threshold was defined as the one that minimises the proportion of misclassified patients; it depends then on the sensitivity and specificity of the diagnostic test [13]. Indeed, more or less weight can be given to these two test characteristics depending on the clinician's preferences and the aversion of the patient toward, on the one hand, local persistent PCa, and on the other hand, biopsy. This methodology was described for other diseases [14, 15]. A very high-risk-averse man might opt for biopsy despite a mere 10% risk of local persistent PCa, and hence, sensitivity will be favoured. Another less risk-averse but a little more anxious about biopsy or an elderly patient might choose a 30 to 40% risk of local persistent PCa before accepting biopsy; in this case, specificity will be favoured. Hence, different PSA nadir thresholds have been defined, depending on the risk of local persistent PCa (hereafter denoted *r*) above which a clinician will propose or a patient will accept biopsy [16]. Two extreme situations have been also defined: i) the patient is too risk-averse; thus, biopsy should be always performed whatever the PSA nadir value; ii) patient is too anxious about biopsy procedure; thus, biopsy should never be performed. Thus, performing a biopsy in case of PSA nadir above a given threshold is considered useful only in patients not too risk-averse and not too anxious about the procedure.

The proportion of positive biopsies among patients whose nadir is above or below the nadir threshold as well as the proportion of patients whose nadir is above the nadir threshold in patients with positive or negative biopsies were calculated for each PSA nadir threshold as defined above. The impact of the pre-HIFU PSA level on the nadir threshold was also analysed; results are given only for *r* equals 30%.

In the routine use of PSA nadir to decide whether to perform biopsy, the nadir will not be calculated from the modelled PSA values, but directly from observed PSA measurements

that are subject to wide fluctuations. Moreover, for each patient, the PSA nadir will be obtained from a limited number of measurements, not from the whole patient's follow-up. Hence, the diagnostic accuracies obtained from nadirs directly calculated from PSA measurements taken during the first three mo after HIFU were compared to the ones stemming from modelled PSA data.

Throughout this article, all confidence intervals are 95% confidence intervals.

3. Results

3.1. Patients and follow-up

The baseline characteristics of the 289 patients that met the inclusion criteria are reported in Table 1. The mean (SD) follow-up was 2.4 yr (2.0) (range: 0.2-7.8). Of these patients, 90 had a positive biopsy between 90 d and one yr after HIFU treatment, which corresponds to a prevalence of 31%.

3.2. Nadir and time to nadir

According to the modelled PSA data, among the patients free from local persistent PCa, the mean nadir (median, 1st quartile, 3rd quartile) was 0.81 ng/ml (0.30, 0.10, 0.83) and was reached after a mean time of 51 days (51, 45, 58). Among the patients who had a local persistent PCa, the mean nadir was 1.97 ng/ml (1.04, 0.28, 2.38), and was reached after a mean time of 44 days (42, 38, 49).

3.3. PSA nadir thresholds

The optimal nadir threshold was estimated to be between 0.09 and 6.39 ng/ml, depending on the risk of local persistent PCa above which a clinician would propose a biopsy (Table 2). For a risk below 0.10, it is considered better to perform systematic biopsies whatever the nadir

value, and avoid biopsy for a risk above 0.74. Depending on the nadir threshold, the proportion of patients that would have a biopsy in case of a nadir above the threshold varies from 4 to 81%; the expected proportion of positives would range from 36% to 58%. If biopsies were carried out in patients with a nadir below the threshold, the proportion of positives would range from 11% to 30%.

There were significant variations between the nadir thresholds estimated for different subgroups depending on the pre-HIFU PSA value; the lower the pre-HIFU PSA values, the lower the nadir thresholds (Table 3).

When using the nadir value to decide whether to perform biopsy, the median delay between the time to nadir and an increase in PSA level of 1.2 ng/ml above the nadir in patients with a positive biopsy was 12.9 mo.

When the nadir thresholds were calculated from the nadirs of PSA measurements taken during the first three mo after HIFU, except for very low nadir thresholds, the diagnostic accuracies were quite close to those obtained with modelled PSA values (Table 4).

4. Discussion

This study focused on the PSA nadir as an early diagnostic rule to initiate biopsies to detect local persistent PCa after HIFU treatment. The diagnostic accuracies of the PSA nadir thresholds cannot be compared to the ones obtained with Blana et *al*. rules [5] because: i) these authors considered local and metastatic PCa; ii) they analysed jointly persistent and recurrent PCa. The median delay of 12.9 mo between the time to nadir and an increase in PSA level of 1.2 ng/ml above the nadir suggests that earlier biopsies could be performed using the nadir rule, and hence, earlier salvage treatment could be initiated. Because the PSA value just before a second HIFU treatment is a strong predictor of the success of the treatment, with lower values associated to better prognoses, it may be preferable to use diagnostic rules that

lead to earlier salvage treatments. The same reasoning motivated Zelefsky et *al*. [17] to use a nadir threshold of 1.5 ng/ml after EBRT to evaluate patients for the presence of persistent cancer. In this case, the rules defined by Blana et *al*. are then more useful to detect recurrent PCa one yr after HIFU.

In this study, we did not define a unique PSA nadir threshold, but a set of thresholds, depending on the risk of local persistent PCa above which a clinician would be willing to perform a biopsy. Hence, the clinician will be able to adapt partly his decision according to patient characteristics such as age or the general health status. The proposed thresholds are neither arbitrary nor based on the quartiles of the nadir distribution but are the ones that maximises the net benefit from biopsy in terms of patient's state of health.

Considering a PSA nadir threshold of 0.22 ng/ml, 187 patients would have a biopsy whereas all 289 patients would have one if systematic biopsies were realised; in the former group, 40% would have a positive biopsy. Only 15% of patients would be missed if biopsy were realised only for nadirs above 0.22 ng/ml. These values change according to the risk of local persistent PCa above which a biopsy is proposed, but the expected benefit in triggering biopsies on nadir results is always substantial.

In the present study, biopsy results during the first 90 d have been discarded because they might be unreliable. The local persistence of PCa was defined according to the results of all biopsies performed during patient follow-up, which might compensate for biopsy lack of sensitivity. Concerning the use of post-HIFU prostate biopsies to assess treatment failure, Rouvière et *al*. [8] have shown that biopsies guided by dynamic contrast-enhanced magnetic resonance imaging detect more cancers than routine biopsies, and hence, the rate of positive biopsies triggered on the PSA nadir results might be nowadays higher. Here, the estimation of the optimal nadir thresholds was based solely on patients who had at least one biopsy over three mo after HIFU; this discarded 66 patients. Hence, the optimal nadir thresholds for the

whole population might be slightly lower than the ones found above. Our results are based on nadirs calculated from modelled PSA values over the whole PSA follow-up, whereas in routine use of PSA, the nadir is calculated directly from observed PSA measurements during a short follow-up period. Still, the diagnostic accuracies obtained using PSA nadirs directly calculated from PSA measurements during the first three mo after HIFU were similar to those obtained from modelled PSA data; this fact justifies the use of the proposed nadir thresholds to decide whether to perform biopsy.

The high number of useless biopsies resulting from systematic biopsies after HIFU could be substantially decreased by triggering biopsies on PSA results: during the first year of follow-up, this should be based on the PSA nadir value to detect local persistent PCa; afterwards, Blana et al. rules may be used to detect local recurrent PCa.

5. Conclusions

The results of the present study promote the use of the PSA nadir as an early diagnostic rule for initiation of biopsies after HIFU. This use leads to good proportions of positive biopsies with low levels of missed cases and, above all, to the possibility to carry out early salvage treatments and better outcomes.

References

1. Blana A, Murat FJ, Walter B, Thuroff S, Wieland WF, Chaussy C, et al. First analysis of the long-term results with transrectal HIFU in patients with localised prostate cancer. Eur Urol 2008; 53: 1194-1201.
2. Blana A, Rogenhofer S, Ganzer R, Lunz J-C, Schostak M, Wieland WF, et al. Eight years' experience with high-intensity focused ultrasonography for treatment of localized prostate cancer. Urology 2008; 72: 1329-1333.
3. Rebillard X, Soulie M, Chartier-Kastler E, Davin J-L, Mignard J-P, Moreau J-L, et al. High-intensity focused ultrasound in prostate cancer; a systematic literature review of the French Association of Urology. BJU Int 2008; 101: 1205-1213.
4. Uchida T, Shoji S, Nakano M, Hongo S, Nitta M, Murota A, et al. Transrectal high-intensity focused ultrasound for the treatment of localized prostate cancer: eight-year experience. Int J Urol 2009; 16: 881-6.
5. Blana A, Brown SC, Chaussy C, Conti GN, Eastham JA, Ganzer R, et al. High-intensity focused ultrasound for prostate cancer: comparative definitions of biochemical failure. BJU Int 2009; 104: 1058-62.
6. Poissonnier L, Chapelon J-Y, Rouvière O, Curiel L, Bouvier R, Martin X, et al. Control of prostate cancer by transrectal HIFU in 227 patients. Eur Urol 2007; 51: 381-387.
7. Pasticier G, Chapet O, Badet L, Ardiet JM, Poissonnier L, Murat FJ, et al. Salvage radiotherapy after high-intensity focused ultrasound for localized prostate cancer: early clinical results. Urology 2008; 72: 1305-1309.
8. Rouviere O, Girouin N, Glas L, Ben Cheikh A, Gelet A, Mege-Lechevallier F, et al. Prostate cancer transrectal HIFU ablation: detection of local recurrences using T2-weighted and dynamic contrast-enhanced MRI. Eur Radiol 2009.
9. Uchida T, Illing RO, Cathcart PJ, Emberton M. To what extent does the prostate-specific antigen nadir predict subsequent treatment failure after transrectal high-intensity focused ultrasound therapy for presumed localized adenocarcinoma of the prostate? BJU Int 2006; 98: 537-539.
10. Ganzer R, Rogenhofer S, Walter B, Lunz J-C, Schostak M, Wieland WF, et al. PSA nadir is a significant predictor of treatment failure after high-intensity focussed ultrasound (HIFU) treatment of localised prostate cancer. Eur Urol 2008; 53: 547-553.
11. Subtil F, Rabilloud M. Robust non-linear mixed modelling of longitudinal PSA levels after prostate cancer treatment. Stat Med 2009; (in press).
12. Soletormos G, Semjonow A, Sibley PEC, Lamerz R, Petersen PH, Albrecht W, et al. Biological Variation of Total Prostate-Specific Antigen: A Survey of Published Estimates and Consequences for Clinical Practice. Clin Chem 2005; 51: 1342-1351.
13. Subtil F, Rabilloud M. Bayesian Method to Estimate the Optimal Threshold of a Longitudinal Biomarker and its Confidence Interval. Biom J 2009; submitted.
14. DeNeef P, Kent DL. Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. Med Decis Making 1993; 13: 126-132.
15. Jund J, Rabilloud M, Wallon M, Ecochard R. Methods to estimate the optimal threshold for normally or log-normally distributed biological tests. Med Decis Making 2005; 25: 406-415.
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26: 565-574.

263 17. Zelefsky MJ, Shi W, Yamada Y, Kollmeier MA, Cox B, Park J, et al.
264 Postradiotherapy 2-Year Prostate-Specific Antigen Nadir as a Predictor of Long-Term
265 Prostate Cancer Mortality. Int J Radiat Oncol Biol Phys 2009.
266
267

Table 1: Baseline characteristics of the 289 patients treated with high-intensity focused ultrasound for prostate cancer.

Characteristic	Value or Number (%)
Mean (median) age at HIFU	69.7 ± 5.6 (71) yr
Mean (median) pre-HIFU PSA	8.25 ± 6.91 (7.15) ng/ml
Clinical stage	
T1	156 patients (54%)
T2	124 patients (43%)
T3	9 patients (3%)
Gleason score	
undefined	5 patients (2%)
≤6	164 patients (57%)
=7	95 patients (33%)
≥8	25 patients (8%)
Mean (median) prostate volume (TRUS)	27.73 ± 13.62 (27) ml

HIFU: high-intensity focused ultrasound - PSA: prostate specific antigen - TRUS: transrectal ultrasound.

Table 2: Optimal nadir threshold estimates [95% confidence interval] according to the risk of local persistent prostate cancer above which a patient would opt for a biopsy, along with the corresponding estimates [95% confidence interval] of diagnostic accuracy.

r	Nadir threshold	Number (%) of prospective biopsies if nadir > threshold [§]	% nadirs > threshold in patients with positive biopsy	% of nadirs > threshold in patients with negative biopsy	% patients with positive biopsies if nadir > threshold	% patients with positive biopsies if nadir ≤ threshold
0.7	6.39 [2.31, 20.34]	13 (0.04)	0.08 [0.05, 0.12]	0.03 [0.01, 0.05]	0.58 [0.37, 0.79]	0.30 [0.71, 0.69]
0.6	3.04 [1.41, 8.11]	31 (0.11)	0.19 [0.14, 0.25]	0.07 [0.04, 0.11]	0.55 [0.41, 0.71]	0.28 [0.73, 0.70]
0.5	1.50 [0.88, 2.94]	61 (0.21)	0.35 [0.28, 0.42]	0.15 [0.10, 0.21]	0.51 [0.41, 0.62]	0.26 [0.77, 0.72]
0.4	0.78 [0.54, 1.18]	100 (0.35)	0.53 [0.45, 0.61]	0.27 [0.20, 0.34]	0.47 [0.40, 0.55]	0.23 [0.81, 0.74]
0.3	0.42 [0.32, 0.56]	142 (0.49)	0.69 [0.61, 0.76]	0.40 [0.33, 0.48]	0.44 [0.38, 0.49]	0.19 [0.85, 0.77]
0.2	0.22 [0.14, 0.30]	187 (0.65)	0.83 [0.76, 0.89]	0.57 [0.49, 0.65]	0.40 [0.36, 0.44]	0.15 [0.90, 0.79]
0.1	0.09 [0.03, 0.16]	235 (0.81)	0.94 [0.89, 0.97]	0.76 [0.68, 0.83]	0.36 [0.34, 0.38]	0.11 [0.95, 0.82]

r: risk of local persistent prostate cancer above which a patient would opt for a biopsy;

[§] number of patients (among 289) that would have a biopsy if biopsy were realised only when the PSA nadir is above the threshold.

Table 3: Optimal nadir threshold estimates [95% confidence interval] when the risk of local persistent prostate cancer above which a patient would opt for a biopsy equals 30%, depending on the pre-HIFU PSA value, along with the corresponding estimates [95% confidence interval] of diagnostic accuracy.

Pre-HIFU PSA value	Nadir threshold	% nadirs > threshold in patients with positive biopsy	% of nadirs > threshold in patients with negative biopsy	% patients with positive biopsies if nadir > threshold	% patients with positive biopsies if nadir ≤ threshold
<4 ng/ml	0.21 [0.08, 0.45]	0.58 [0.38, 0.76]	0.33 [0.21, 0.46]	0.58 [0.44, 0.69]	0.32 [0.21, 0.43]
≥4 to <10	0.46 [0.30, 0.69]	0.71 [0.56, 0.86]	0.38 [0.30, 0.48]	0.63 [0.55, 0.70]	0.30 [0.17, 0.41]
≥10 to <20	0.80 [0.49, 1.19]	0.75 [0.59, 0.89]	0.36 [0.24, 0.49]	0.83 [0.77, 0.88]	0.46 [0.28, 0.61]

HIFU: high-intensity focused ultrasound; PSA: prostate specific antigen.

284 **Table 4**: Estimates of the diagnostic accuracy [95% confidence interval] for several PSA nadir thresholds when the nadir is calculated directly
285 from observed PSA measurements taken during the first three mo following HIFU.

Optimal nadir threshold	% nadirs > threshold in patients with positive biopsy	% of nadirs > threshold in patients with negative biopsy	% patients with positive biopsies if nadir > threshold	% patients with positive biopsies if nadir ≤ threshold
6.39	0.09 [0.04, 0.17]	0.03 [0.01, 0.06]	0.57 [0.29, 0.82]	0.30 [0.24, 0.36]
3.04	0.21 [0.13, 0.31]	0.08 [0.05, 0.13]	0.53 [0.35, 0.70]	0.28 [0.23, 0.34]
1.5	0.41 [0.31, 0.52]	0.16 [0.13, 0.24]	0.51 [0.39, 0.63]	0.24 [0.19, 0.31]
0.78	0.56 [0.45, 0.66]	0.23 [0.22, 0.34]	0.48 [0.38, 0.58]	0.22 [0.16, 0.28]
0.42	0.64 [0.54, 0.74]	0.37 [0.36, 0.50]	0.41 [0.32, 0.49]	0.22 [0.16, 0.30]
0.22	0.80 [0.70, 0.88]	0.48 [0.46, 0.60]	0.40 [0.33, 0.48]	0.16 [0.10, 0.24]
0.09	0.90 [0.82, 0.95]	0.71 [0.67, 0.79]	0.36 [0.29, 0.42]	0.15 [0.07, 0.26]

286

6.2.3.2 Principaux résultats

Des seuils de nadir de PSA ont été définis pour des risques de maladie pour la réalisation de biopsies compris entre 11 et 70 % ; en dessous, il est préférable de réaliser systématiquement une biopsie ; au dessus, la réalisation d'une biopsie est exclue. Les valeurs prédictives positives des différents seuils de nadirs proposés – c'est à dire le pourcentage de biopsies positives attendues chez les patients dont la valeur de nadir est supérieure au seuil – variait entre 36 et 58 %. Les valeurs prédictives négatives variaient entre 70 et 89 %, signifiant que si des biopsies étaient réalisées pour les patients ayant un nadir en dessous du seuil, 11 à 30 % se révéleraient positives. Il faut, encore une fois, rappeler que ces estimations dépendent totalement de la prévalence de la population étudiée ; elles ne sont pas à transposer directement à d'autres populations où la prévalence est différente.

Les seuils de PSA ainsi définis conduiraient à la réalisation de biopsies pour 8 à 81 % des patients ; même une diminution de 20 % du nombre de biopsies réalisées par rapport à la stratégie de biopsie systématique est non négligeable.

Les performances diagnostiques obtenues à partir des nadirs de PSA modélisés et de ceux qui sont mesurés directement durant les trois premiers mois de suivi étaient assez similaires, indiquant que les seuils de nadirs définis sont également applicables pour des nadirs cliniques.

6.2.4 Compléments à l'article

6.2.4.1 Courbes de décision

La figure 6.1 représente les courbes du bénéfice espéré de la réalisation d'une biopsie pour les trois stratégies possibles :

- stratégie 1 : déclencher la biopsie en fonction du nadir de PSA ;
- stratégie 2 : ne jamais faire de biopsie ;
- stratégie 3 : réaliser systématiquement une biopsie.

Les deux traits verticaux indiquent les valeurs moyennes de risque de maladie pour la réalisation de biopsies pour lesquelles la stratégie liée à l'utilisation des PSA a un bénéfice espéré non statistiquement différent de celui des deux autres stratégies.

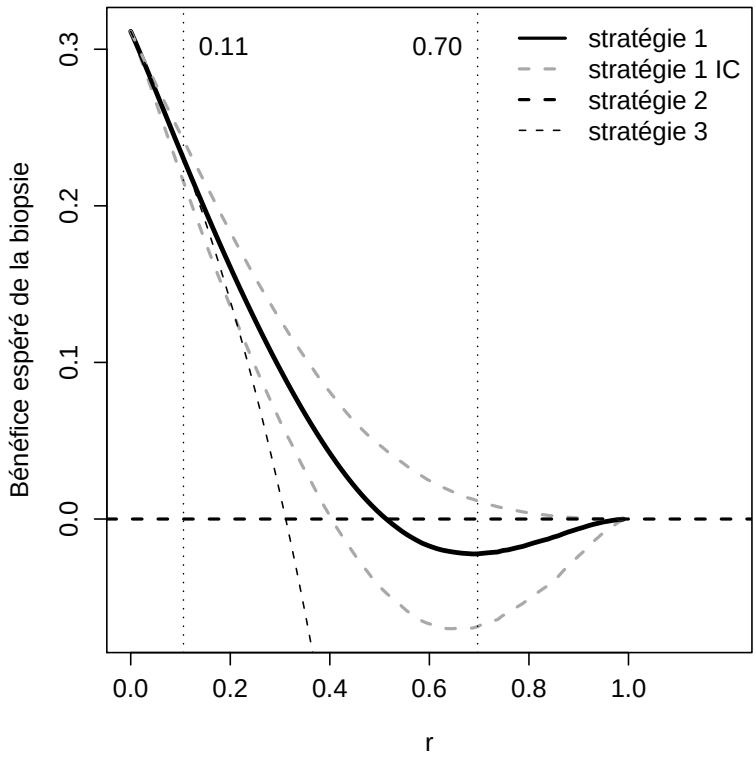

Figure 6.1 – Courbes du bénéfice espéré de la réalisation d'une biopsie pour trois stratégies différentes (IC : intervalle de crédibilité).

6.2.4.2 Validation croisée

Dans le travail effectué précédemment, les performances diagnostiques liées à l'utilisation des différents seuils de nadir de PSA pour discriminer les patients ont été estimées sur les mêmes données que celles ayant permis d'établir ces seuils. Il se peut, par conséquent, que les performances estimées surestiment celles qui auraient été obtenues avec les mêmes seuils, mais dans une population différente. Ceci correspond au problème de l'optimisme des modèles (Gerds et *al.*, 2008).

La validation croisée peut permettre de corriger l'optimisme des modèles (Efron et Tibshirani, 1997). Cette méthode consiste à diviser le jeu de données en deux parties : la première permet l'estimation du seuil, la seconde l'estimation des performances diagnostiques obtenues en retenant cette valeur de seuil. Cette opération est répétée plusieurs fois ; les résultats obtenus sont ensuite moyennés. De cette façon, les performances diagnostiques ne sont pas estimées sur les mêmes données que celles ayant servi à déterminer les seuils.

La méthode des K partitions est courante en validation croisée. Elle consiste a subdiviser le jeu de données en K parties, à estimer le seuil sur toutes les parties sauf la k$^{\text{ième}}$, puis à évaluer

les performances sur cette partie. L'opération est répétée en changeant la partie sur laquelle les performances sont évaluées, les résultats des différentes parties étant ensuite moyennés. Dans le cas de l'évaluation du nadir de PSA, les parties ont été constituées de telle sorte que la proportion de malades et non malades soit conservée dans chaque partie par rapport à la population initiale.

Le tableau 6.1 présente les performances diagnostiques estimées par validation croisée pour les différents niveaux de risque de maladie à partir desquels une biopsie est proposée. Globalement, les résultats sont très similaires à ceux obtenus sans validation croisée ; ainsi, la surestimation des performances diagnostiques était limitée.

Tableau 6.1 – Performances diagnostiques estimées avec validation croisée pour les différents risques de maladie pour la réalisation de biopsie.

r	Sen	Spe	Vpp	Vpn
0,7	0,09 (0,08)	0,96 (0,97)	0,55 (0,58)	0,70 (0,70)
0,6	0,22 (0,19)	0,92 (0,93)	0,57 (0,55)	0,72 (0,72)
0,5	0,39 (0,35)	0,83 (0,85)	0,51 (0,51)	0,75 (0,74)
0,4	0,56 (0,53)	0,73 (0,73)	0,49 (0,47)	0,79 (0,77)
0,3	0,63 (0,69)	0,58 (0,60)	0,41 (0,44)	0,78 (0,81)
0,2	0,82 (0,83)	0,46 (0,43)	0,41 (0,40)	0,85 (0,85)
0,1	0,90 (0,94)	0,28 (0,24)	0,36 (0,36)	0,86 (0,89)

() : performances estimées sans validation croisée.

6.2.5 Patients ayant moins de cinq mesures de PSA

Au cours de l'étude, les patients ayant moins de cinq mesures ont été retirés de l'analyse pour des raisons techniques (les chaînes MCMC des effets aléatoires de ces patients ne convergeaient pas en raison du nombre trop faible de mesures). Dans l'article destiné à The European Urology, les sensibilités et spécificités associées aux différents seuils de nadir de PSA définis ont été recalculées à partir des valeurs de nadir de PSA réellement observées durant les trois premiers mois de suivi. Cette même analyse est effectuée ici, mais en incluant les patients ayant moins de cinq mesures de PSA au total (tableau 6.2).

Les patients avec moins de cinq mesures de PSA avaient en général des mesures plus espacées dans le temps que ceux ayant plus de cinq mesures ; sur l'ensemble des patients ayant

Tableau 6.2 – Performances diagnostiques estimées à partir des valeurs de nadir de PSA obtenues grâces aux mesures de PSA durant les trois premiers mois de suivi, pour les patients ayant plus et moins de cinq mesures de PSA.

r	Sen	Spe
0,7	0,13 (0,09)	0,95 (0,97)
0,6	0,25 (0,21)	0,90 (0,92)
0,5	0,44 (0,41)	0,82 (0,84)
0,4	0,59 (0,56)	0,72 (0,77)
0,3	0,69 (0,64)	0,58 (0,63)
0,2	0,82 (0,80)	0,47 (0,52)
0,1	0,92 (0,90)	0,28 (0,29)

() : performances estimées pour les patients ayant plus de cinq mesures.

plus de 90 jours de suivi, les patients avec moins de cinq mesures peuvent être considérés comme des patients ayant manqué des mesures. Il faut rappeler que les valeurs de PSA d'un patient diminuent jusqu'au nadir, puis ré-augmentent. Ainsi, avec moins de mesures de PSA, le nadir de PSA estimé risque d'être plus élevé, en moyenne, pour les patients ayant moins de cinq mesures. Ceci est confirmé dans le tableau 6.3 concernant les valeurs de nadir observées.

Ainsi, les sensibilités obtenues sur l'ensemble des patients doivent être plus élevées, et les spécificités plus faibles. D'après le tableau 6.2, les augmentations de sensibilité sont modestes ; de même, les diminutions de spécificité sont peu élevées. Globalement, la sélection des patients ayant plus de cinq mesures de PSA – pour des raisons techniques – a entraîné une modification de la population étudiée ; néanmoins, les valeurs de performances diagnostiques estimées n'en sont pas trop modifiées.

6.3 Bilan du chapitre 6

Dans cette partie, plusieurs seuils de nadir de PSA ont été établis en fonction des préférences des patients et des cliniciens en termes de sensibilité et de spécificité ou de qualité et de quantité de vie, caractérisées, dans les deux cas, par le risque de maladie à partir duquel une biopsie est proposée ou acceptée. Ces résultats sont valables pour des patients ayant plus de 90 jours de suivi pour lesquels il y a déjà une suspicion de persistance de cellules cancéreuses. Des

Tableau 6.3 – Quartiles et moyennes des valeurs de nadirs de PSA observées durant les trois premiers mois de suivi chez les patients malades et non malades, en distinguant les patients ayant plus de cinq mesures de PSA de ceux ayant moins de cinq mesures.

	1er quartile	2nd quartile	Moyenne	3ème quartile
Malades				
Moins de 5 mesures	0,38	1,27	3,96	3,70
Plus de 5 mesures	0,31	0,99	1,87	1,99
Non malades				
Moins de 5 mesures	0,09	0,32	2,02	0,90
Plus de 5 mesures	0,08	0,25	0,90	0,98

limites basses et hautes de risques ont été définies afin de ne pas estimer des seuils de nadir dont l'utilisation pour déclencher les biopsies ne conduirait pas à un bénéfice espéré meilleur que celui des stratégies " biopsie systématique " ou aucune biopsie.

La détermination de plusieurs seuils permet d'individualiser la prise de décision en fonction des caractéristiques du patient, par exemple selon l'état de santé général ou l'âge. Elle permet également de tenir compte des préférences du patient en termes de qualité de vie ou de quantité de vie, et ce, au travers du risque de maladie justifiant la réalisation d'une biopsie. L'article rédigé pour The European Urology montre également que, suivant les valeurs de PSA pré traitement, le seuil estimé dans les sous groupes peut varier, celui-ci étant d'autant plus élevé que le taux de PSA pré traitement est élevé. L'inconvénient de cette approche est qu'elle réduit à chaque fois la population d'étude ; les intervalles de crédibilité des seuils optimaux s'en trouvent ainsi augmentés. Une solution consisterait à modéliser la distribution des seuils de nadirs dans les groupes malades et non malades en fonction des caractéristiques des patients. Ainsi, l'ensemble de la population de départ participerait à l'estimation de tous les seuils, quels que soient les niveaux considérés des caractéristiques.

D'un point de vue clinique, le déclenchement de biopsie en fonction du nadir de PSA semble être une meilleur option que la stratégie de biopsie systématique pour les patients pour lesquels il y a déjà une suspicion de persistance de cellules cancéreuses. Il permet de limiter grandement le nombre de biopsies réalisées, en manquant peu de patients (patients au nadir inférieur au seuil mais pour lesquels une biopsie se serait tout de même révélée positive). De plus, le nadir étant atteint durant les trois premiers mois pour la plupart des patients, cette stratégie permet de déclencher rapidement un traitement de sauvetage. Les seuils de nadir de

PSA proposés sont utiles pour détecter la persistance locale de cellules cancéreuses, mais pas forcément pour les patients qui récidivent après un an de suivi. Pour ces patients, d'autres tests plus tardifs ont été proposés par Blana et *al.* (2009).

Dans ce chapitre, il a été défini plusieurs seuils de nadir de PSA tenant compte des préférences des médecins ou des patients. Dans les deux cas, ces préférences sont caractérisées par le risque de maladie justifiant la réalisation d'une biopsie, ou par le ratio bénéfice net sur coût net. Le résultat final résulte en général d'un consensus entre le médecin et son patient, mais il n'a jamais été évoqué le fait que les avis des médecins et des patients peuvent être divergents, rendant cette méthode difficilement applicable. La décision finale devrait revenir au patient, éclairé par l'avis du médecin, mais cette réflexion déborde du cadre de cette thèse.

Quatrième partie

Perspectives

Chapitre

7 Prise en compte du gold standard imparfait

Dans le cadre des données de PSA après UFHI, le statut des patients par rapport à la persistance de cellules cancéreuses a été défini à partir du résultat des biopsies de prostate. En réalité, la biopsie n'est qu'un proxy imparfait de la persistance de cellules cancéreuses : une biopsie n'est positive que si des cellules cancéreuses subsistent localement, mais l'inverse n'est pas vrai. Une biopsie peut être négative car les prélèvements ont été effectués dans une zone de la prostate ne présentant pas de cellules cancéreuses. La spécificité des biopsies est donc de 100 %, mais ce n'est pas le cas de la sensibilité ; le résultat combiné de plusieurs biopsies pour un même patient doit toutefois permettre d'améliorer la sensibilité globale. De plus, les techniques modernes d'imagerie médicale permettent de focaliser les biopsies dans les zones à risque (Rouvière et *al.*, 2010). Les biopsies sont néanmoins à considérer comme un *gold standard imparfait* pour le diagnostic de persistance de cellules cancéreuses ; ceci peut entraîner une sous-estimation de la spécificité du nadir de PSA, ainsi qu'une surestimation de sa sensibilité. Le terme de gold standard imparfait est souvent employé dans la littérature et le sera dans la suite de ce chapitre ; en réalité, il ne fait que décrire les situations dans lesquelles il n'y a pas de gold standard.

Dans les chapitres précédents, pour contourner le problème du gold standard imparfait, le nadir de PSA a été analysé comme un test permettant de discriminer les patients suivant le résultat attendu des biopsies et non selon la persistance de cellules cancéreuses. Ainsi, les sensibilités et spécificités estimées sont justes par rapport à l'objectif " correspondre au résultat des biopsies ". Pour autant, le souhait des cliniciens pourrait être de se passer totalement des

biopsies, la décision de retraitement étant alors prise uniquement en fonction du nadir de PSA obtenu durant les trois premiers mois de suivi. Il serait nécessaire, dans ce cas, de corriger les estimations de sensibilité et de spécificité. Une approche permettant de tenir compte du fait que la biopsie correspond à un gold standard imparfait est présentée par la suite.

7.1 Estimation de performances diagnostiques en situation de gold standard imparfait

Le résultat croisé de deux tests diagnostiques imparfaits sur une population fournit trois degrés de liberté, alors que cinq paramètres sont à estimer : les sensibilités et spécificités des deux tests ainsi que la prévalence dans la population étudiée. Si au moins deux des paramètres sont connus, les trois restant sont estimables par maximum de vraisemblance. Dans le cas contraire, le nombre de degrés de liberté peut être augmenté en recherchant le résultat croisé des deux tests sur une autre population dans laquelle la prévalence est différente, ou en combinant le résultat d'un troisième test imparfait.

Lorsque l'on ne dispose des résultats que d'un seul test, une solution consiste à définir une variable latente correspondant au statut inconnu des patients vis à vis de la maladie, ainsi qu'à introduire de l'information a priori sur les différents paramètres (sensibilité, spécificité et prévalence) à partir d'études antérieures (Joseph et *al.*, 1995 ; Choi et *al.*, 2006). Conditionnellement au statut latent des patients, il est possible d'échantillonner dans la distribution a posteriori de la sensibilité, de la spécificité et de la prévalence. De façon similaire, conditionnellement à ces trois paramètres, la distribution a posteriori du statut des patients est connue. L'échantillonneur de Gibbs est particulièrement utile dans ce type d'approche où il est nécessaire d'échantillonner dans les distributions conditionnelles.

Parfois, l'objectif porte plus sur l'estimation de la prévalence en situation de gold standard imparfait. Un certain nombre de covariables peuvent renseigner sur la valeur de celle-ci au travers, par exemple, d'une régression logistique. En introduisant de l'information a priori sur l'effet de ces covariables sur la prévalence, il est possible d'obtenir une estimation corrigée de la prévalence (Tu et *al.*, 1999). Pour ce scénario, McInturff et *al.* (2004) ont proposé une méthode qui évite l'introduction dans le modèle d'une variable latente correspondant au statut des patients.

Lorsque les résultats de deux tests sont disponibles, Black et Craig (2002) ont proposé d'en modéliser les résultats conjoints, en introduisant de l'information a priori sur les performances des deux tests et sur la prévalence. La difficulté est de modéliser correctement la relation de

dépendance entre les résultats des tests conditionnellement au statut latent des patients. Une hypothèse simplificatrice est de considérer que les résultats des deux tests sont conditionnellement indépendants (Joseph et *al.*, 1995), mais ceci n'est pas toujours vrai. Supposons, de manière imaginaire, que le nombre de cellules prostatiques puisse être utilisé comme un test diagnostique du cancer de la prostate. Les résultats des tests basés sur les PSA et sur le nombre de cellules prostatiques ne sont pas indépendants conditionnellement au vrai statut du patient car, indépendamment de la présence de cancer, le taux de PSA est relié au nombre de cellules prostatiques. Black et Craig (2002) ont proposé de construire plusieurs modèles supposant différentes relations de dépendance entre les tests, puis d'estimer les performances diagnostiques de chacun des tests en moyennant les résultats obtenus selon les différents modèles par Bayesian Model Averaging (Hoeting et *al.*, 1999).

Cette dernière méthode reste complexe. Dans le cadre des données de PSA, une solution peut consister à ne pas utiliser l'information concernant le nadir des PSA dans la détermination du statut latent des patients, mais uniquement le résultat des biopsies et d'autres facteurs prédisposant à un éventuel échec du traitement UFHI. Le statut latent ainsi obtenu sert de gold standard pour l'estimation des performances du nadir de PSA ; de cette façon, la relation entre le nadir et le résultat des biopsies n'a pas à être modélisée.

7.2 Nadir de PSA et diagnostic de persistance de cellules cancéreuses

7.2.1 Principe

Un certain nombre de covariables au moment du premier traitement UFHI peuvent influencer la probabilité d'échec du traitement, ou de persistance de cellules cancéreuses. Les patients ayant des niveaux de PSA pré traitement élevés ou des stades cliniques de cancer élevés sont en général plus prédisposés à un échec du traitement. De la même façon, la localisation géographique de la tumeur peut fortement influencer la probabilité d'échec du traitement : un cancer se situant proche de l'apex a moins de chance d'être guéri, les tirs d'ultrasons étant focalisés en dehors de cette zone afin d'épargner l'apex.

De manière générale, on suppose qu'il existe K covariables reliées au risque d'échec de traitement et que cette relation est modélisable au travers d'une régression logistique. Soit t_i la probabilité de persistance de cellules cancéreuses pour le patient i, X_{ik} la valeur de la $k^{\text{ième}}$

covariable pour ce patient et β_k l'effet de cette covariable sur la probabilité de persistance de cancer. Le statut latent, noté z_i, suit une loi de Bernoulli de paramètre p_i. Le résultat des biopsies est fonction de ce statut, ainsi que de leur sensibilité et spécificité. Les relations entre les différentes variables sont représentées sur la figure 7.1. b_{ij} correspond au résultat de la $j^{\text{ième}}$ biopsie du $i^{\text{ème}}$ patient ayant u_i biopsies ; η dénote par la suite la sensibilité de la biopsie et ϑ la spécificité (fixée à 1). En présence d'information a priori à propos de la valeur de la sensibilité, ainsi que de l'effet des covariables sur la probabilité de persistance de cancer, il est possible d'échantillonner des valeurs dans la distribution a posteriori du statut latent des patients. Ces valeurs permettent ensuite l'estimation des performances diagnostiques du nadir de PSA pour discriminer les patients selon la persistance de cellules cancéreuses.

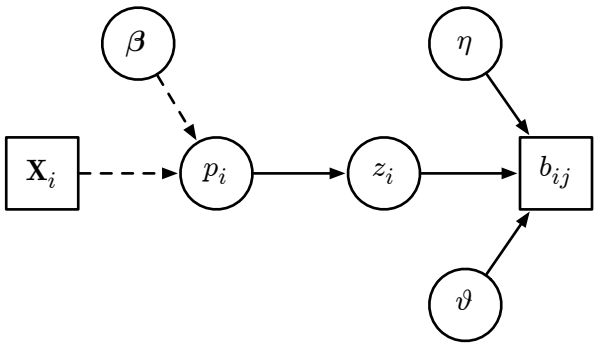

Figure 7.1 – Modèle pour la détermination du statut latent des patients.

7.2.2 Modèle

7.2.2.1 Sensibilité

D'après la figure 7.1, conditionnellement au statut latent des patients, la sensibilité est indépendante des autres paramètres du modèle. La distribution conditionnelle de η est donnée par :

$$P(\eta|\mathbf{b},\mathbf{z}) \propto P(\mathbf{b}|\eta,\mathbf{z})P(\eta)$$

Par la suite, on suppose que les résultats des biopsies d'un même patient sont indépendants conditionnellement au statut latent de ce patient. Cette hypothèse est valide si tout cancer peut être détecté pour chaque patient dès la première biopsie ; elle est moins valide pour les patients qui effectuent une récidive tardivement et pour lesquels l'ordre de la réalisation des biopsies a

une influence sur le résultat. Sous cette hypothèse :

$$P(\mathbf{b}|\eta,\mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{u_i} P(b_{ij}|\eta,\mathbf{z})$$

Lorsque le patient est malade, le résultat de la biopsie est donné par une loi de Bernoulli de paramètre η ; lorsqu'il n'est pas malade, le résultat est également donné par une loi de Bernoulli de paramètre $1-\vartheta = 0$. Ainsi, conditionnellement au statut du patient, le résultat d'une biopsie suit une loi de Bernoulli de paramètre $\eta^{z_i}0^{1-z_i}$. Si une loi beta de paramètre a_η et b_η est utilisée comme a priori pour la distribution du paramètre de sensibilité, la densité conditionnelle est donnée par :

$$P(\eta|\mathbf{b},\mathbf{z}) \propto \prod_{i=1}^{n}\prod_{j=1}^{u_i} \left(\eta^{z_i}0^{1-z_i}\right)^{b_{ij}} \left(1-\eta^{z_i}0^{1-z_i}\right)^{1-b_{ij}} \eta^{a_\eta-1}(1-\eta)^{b_\eta-1}$$

Cette densité ne correspond pas à une loi de probabilité connue, mais étant donné qu'elle est log concave, il est possible d'échantillonner dans la distribution conditionnelle en utilisant un algorithme de type adaptative rejection sampling (Gilks et *al.*, 1996).

7.2.2.2 Effet des covariables sur la probabilité de persistance de cellules cancéreuses

La relation entre la probabilité de persistance de cellules cancéreuses après le traitement et les covariables pré traitement est supposée modélisable au travers d'une régression logistique :

$$\text{logit}(p_i) = \boldsymbol{\beta}\mathbf{X}_i$$

Conditionnellement au statut latent du patient, l'effet des covariables est indépendant des autres paramètres ; la distribution conditionnelle est donc obtenue par :

$$P(\boldsymbol{\beta}|z_i,\mathbf{X}) \propto P(z_i|\boldsymbol{\beta},\mathbf{X})P(\boldsymbol{\beta})$$

Si l'a priori concernant l'effet des covariables est décrit par une loi normale multivariée ($\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$), la densité conditionnelle est donnée par :

$$P(\boldsymbol{\beta}|z_i,\mathbf{X}) \propto \prod_{i=1}^{n} {p_i}^{z_i}(1-p_i)^{1-z_i} \exp\left((\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T{\boldsymbol{\Sigma}_0}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T\right)$$

où $p_i = \exp(\boldsymbol{\beta}\mathbf{X}_i)/(1+\exp(\boldsymbol{\beta}\mathbf{X}_i))$. Encore une fois, il est possible d'échantillonner dans la distribution conditionnelle de $\boldsymbol{\beta}$ par adaptative rejection sampling, car la densité conditionnelle est log concave.

7.2.2.3 Statut latent des patients

La densité conditionnelle du statut des patients est donnée par :

$$\begin{aligned} P(z_i|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{X}_i) &\propto P(\mathbf{b}_i|z_i)P(z_i|\boldsymbol{\beta},\mathbf{X}_i) \\ &\propto \prod_{j=1}^{u_i} \left(\eta^{z_i}0^{1-z_i}\right)^{b_{ij}} \left(1-\eta^{z_i}0^{1-z_i}\right)^{1-b_{ij}} p_i^{z_i}(1-p_i)^{1-z_i} \end{aligned}$$

Ce statut latent est échantillonné dans une loi de Bernoulli de paramètre :

$$\frac{P(z_i=1|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{X}_i)}{(P(z_i=1|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{X}_i)+P(z_i=0|\mathbf{b}_i,\boldsymbol{\beta},\mathbf{X}_i))}$$

L'échantillonneur de Gibbs peut être utilisé pour échantillonner successivement dans les distributions conditionnelles et obtenir ainsi la distribution a posteriori du statut latent de chaque patient.

7.2.3 Estimation des performances du nadir de PSA

Soit nadir_i la valeur du nadir de PSA observé durant les trois premiers mois de suivi pour le patient i. Soit $p_i = \sum_{g=1}^{G} z_{ig}/G$ la moyenne de la distribution a posteriori du statut latent de ce patient, z_{ig} correspondant à la valeur échantillonnée à la $g^{\text{ième}}$ itération de l'algorithme MCMC. Les moyennes des distributions a posteriori de la sensibilité et de la spécificité du nadir de PSA pour un seuil c sont données par :

$$\mu_{\text{Sen}(c)} = \frac{\sum_{i=1}^{n} I(\text{nadir}_i > c) \times p_i}{\sum_{i=1}^{n} p_i} \qquad \mu_{\text{Spe}(c)} = \frac{\sum_{i=1}^{n} I(\text{nadir}_i \leq c) \times (1-p_i)}{\sum_{i=1}^{n} (1-p_i)}$$

Pour la sensibilité, la formule correspond à la moyenne des nadirs supérieurs au seuil c pour toute la population, pondérée par la probabilité de maladie de chacun des patients. L'interprétation de la formule est similaire pour la spécificité.

Un exemple de résultats est proposé en utilisant, comme information a priori à propos de l'effet des covariables sur la probabilité d'échec du traitement, l'information issue des mêmes données que celles de l'étude, en effectuant une régression logistique du résultat des biopsies en

fonction des covariables. Cette approche n'est pas valide, l'information a priori devant provenir des résultats d'une autre étude. Les résultats ci-après sont donc présentés uniquement à titre d'exemple illustratif, pour comprendre la démarche dans sa globalité. Deux covariables ont été retenues pour modéliser la probabilité d'échec de traitement :

- le taux de PSA pré traitement, catégorisé en quatre groupes : inférieur à 4 ng/mL, compris entre 4 et 10, entre 10 et 20 et supérieur à 20 ng/mL ;
- le stade clinique du cancer au moment du premier traitement, avec deux groupes : stade T1 et stades T2-T3.

Pour la sensibilité des biopsies, une loi beta de paramètres 33,38 et 62 a été choisie, correspondant à une sensibilité moyenne a priori de 0,35. 1000 itérations de la chaîne MCMC ont été retenues après une phase de chauffe de 10 000 itérations.

La figure 7.2 correspond à l'histogramme des valeurs moyennes de probabilité de persistance de cancer prédites par le modèle pour chaque patient. Ceux dont la probabilité moyenne est de 1 correspondent aux patients dont une biopsie était positive et pour lesquels il n'y avait pas d'incertitude sur le statut. Pour les patients dont toutes les biopsies étaient négatives, les probabilités de persistance prédites varient entre 40 et 83 %. Ainsi, l'information apportée par les covariables permet de faire varier fortement la probabilité prédite de persitance de cellules cancéreuses. D'après ces résultats, il est probable qu'un certain nombre de patients ayant été classés, d'après les résultats des biopsies, comme étant totalement guéris du cancer avaient encore des cellules cancéreuses.

Le tableau 7.1 représente les performances du nadir de PSA observé durant les trois premiers mois de suivi pour discriminer les patients selon la persistance de cellules cancéreuses. Les résultats sont fournis pour les différents seuils de nadir définis dans le chapitre six selon le risque de maladie à partir duquel une biopsie est conseillée. Globalement, les valeurs de sensibilité obtenues sont plus faibles que celles obtenues en supposant la biopsie constituer un gold standard parfait ; les valeurs de spécificité ne changent quant à elles pas beaucoup. Ces résultats vont dans le sens attendu par un défaut de sensibilité des biopsies. Néanmoins, il ne faut pas interpréter ces valeurs, puisque les a priori utilisés ne sont pas corrects.

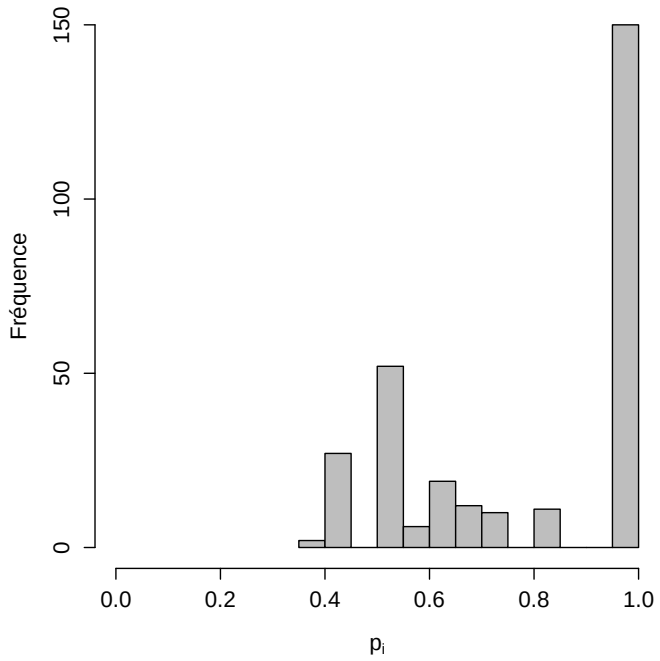

Figure 7.2 – Histogramme des valeurs moyennes de probabilité prédite de persistance de cellules cancéreuses.

7.2.4 Poids de l'information a priori

Les résultats obtenus à partir du modèle précédent dépendent très fortement de l'information a priori utilisée. Une analyse de sensibilité quant au choix de cette information a priori est donc conseillée. La méthode des *power priors* (Ibrahim et Chen, 2000) permet la réalisation d'une telle analyse. Le principe consiste à donner plus ou moins de poids à l'information a priori dans la distribution a posteriori. En supposant que l'information a priori sur les paramètres $\boldsymbol{\beta}$ ait été obtenue à partir d'un jeu de données D_0, elle peut être spécifiée sous la forme :

$$P(\boldsymbol{\beta}|D_0, a_0) \propto (P(D_0|\boldsymbol{\beta}))^{a_0} P_0(\boldsymbol{\beta})$$

$P(\boldsymbol{\beta}|D_0, a_0)$ dénote la probabilité a priori de $\boldsymbol{\beta}$ d'après le jeu de données antérieur D_0 ; $P(D_0|\boldsymbol{\beta})$ correspond à la vraisemblance des données D_0 sachant $\boldsymbol{\beta}$ et $P_0(\boldsymbol{\beta})$ à la probabilité a priori lors de l'estimation des paramètres à partir des données D_0. a_0 est un paramètre de dispersion contrôlant les extrémités de la distribution a priori pour $\boldsymbol{\beta}$. Plus a_0 est proche de 0, plus la distribution a posteriori fournie par les données D_0 est large ; elle a donc moins d'impact sur la distribution a posteriori.

Tableau 7.1 – Performances diagnostiques du nadir de PSA pour discriminer les patients selon la persistance de cellules cancéreuses.

r	Sen	Spe	Vpp	Vpn
0,7	0,28 (0,35)	0,87 (0,86)	0,89 (0,74)	0,24 (0,55)
0,6	0,46 (0,53)	0,71 (0,68)	0,86 (0,64)	0,25 (0,57)
0,5	0,56 (0,64)	0,61 (0,60)	0,85 (0,63)	0,27 (0,61)
0,4	0,69 (0,79)	0,52 (0,50)	0,85 (0,63)	0,31 (0,68)
0,3	0,76 (0,83)	0,39 (0,39)	0,83 (0,60)	0,30 (0,68)
0,2	0,83 (0,88)	0,30 (0,29)	0,82 (0,57)	0,31 (0,69)

() : performances obtenues en considérant les biopsies comme un gold standard imparfait.

Le fait de pondérer la distribution a priori est important lorqu'il existe une hétérogénéité entre les données servant à construire l'a priori et les données de l'étude considérée, ou lorsque les tailles des deux jeux de données sont très différentes. L'analyse de sensibilité peut être menée en faisant varier la valeur de a_0, puis en mesurant l'impact sur les valeurs de performances diagnostiques estimées.

7.3 Perspectives

Un objectif futur serait d'acquérir les données d'un autre centre hospitalier réalisant des traitements UFHI du cancer de la prostate, par exemple celui de Regensburg en Allemagne, afin de déterminer la distribution a priori des covariables pouvant influencer la probabilité de persistance du cancer. Une partie importante du travail porterait sur le choix des covariables à inclure dans le modèle, ainsi que sur l'impact de ce choix sur les résultats estimés en termes de performances diagnostiques. Le Bayesian Model Averaging (Hoeting et *al.*, 1999) permettrait, dans ce cas, de moyenner les résultats issus des différents modèles constructibles, et ainsi, de tenir compte de l'incertitude quant au choix de ces modèles dans l'estimation des performances. Les résultats fournis seraient ainsi plus prudents.

Chapitre

8 Intégration du temps dans l'estimation du seuil

A l'issu du traitement UFHI, parmi les patients ayant une biopsie positive (150), la biopsie est positive durant la première année de suivi pour plus de la moitié des patients (90) ; les dates de biopsie positive des patients restants sont plus dispersées et varient entre un et six ans (figure 8.1). Au cours du traitement UFHI, certaines régions de la prostate peuvent ne pas être traitées, entre autres afin de préserver les bandelettes neurovasculaires, ou parce qu'elles sont difficilement accessibles. Si des cellules cancéreuses étaient présentes dans ces régions, alors elles seront détectées par biopsie immédiatement après le traitement (si toutefois le prélèvement est effectué dans ces zones). Pour d'autres patients, il peut persister après le traitement des cellules cancéreuses, mais qui ne sont pas encore actives ; la vitesse d'évolution de ces cellules est hétérogène d'un patient à l'autre. Ces patients sont en phase préclinique, le cancer n'est pas symptomatique ; les biopsies réalisées durant cette phase sont négatives.

Ainsi, il existe une hétérogénéité de l'évolution de la maladie entre les patients, qui se traduit par une hétérogénéité dans les dates de biopsie positive, mais également dans les valeurs de nadirs de PSA. Pour l'analyse, trois périodes de positivation de la biopsie sont définies :

- biopsie positive durant les 170 premiers jours suivant le traitement (période 1) ;
- biopsie positive entre 170 et 480 jours (période 2) ;
- biopsie positive au delà de 480 jours (période 3).

Ces découpages ont été effectués de telle sorte qu'il y ait à peu près 50 patients malades au sein de chaque période. La figure 8.2 représente les valeurs de logarithme de nadir de PSA en fonction de la période de positivation de la biopsie, les patients n'ayant pas eu de biopsie positive

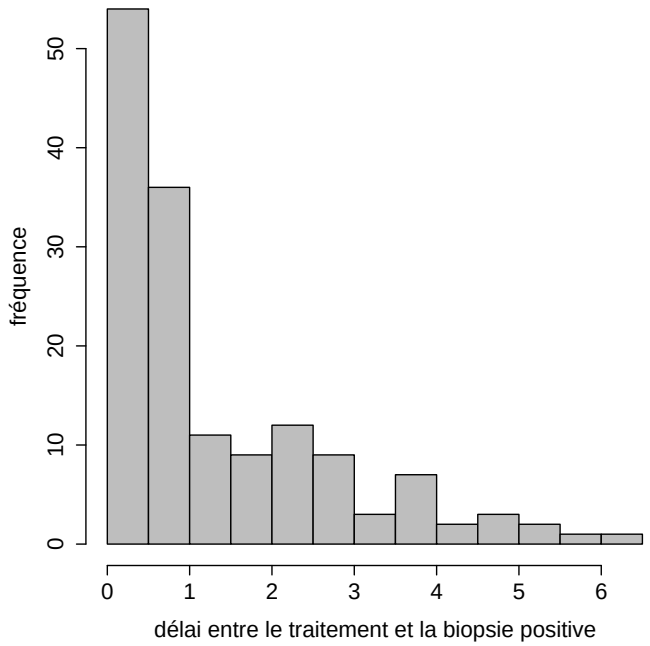

Figure 8.1 – Histogramme des délais en années entre le traitement UFHI et la biopsie positive, pour les patients ayant eu une biopsie positive.

étant classés dans la catégorie " biopsies négatives ". Les valeurs de nadir sont dans l'ensemble plus élevées pour les patients dont la persistance de cellules cancéreuses est détectée durant la première période que durant la deuxième période, et de même entre la deuxième période et la troisième. Ainsi, plutôt que de définir un unique seuil de nadir de PSA pour déclencher les biopsies, il peut être plus judicieux d'en estimer un pour chacune des périodes, et de définir la date de réalisation de la biopsie en fonction de la période au cours de laquelle elle risque d'être positive.

Dans l'article rédigé pour The European Urology, le suivi des patients a été découpé en deux périodes ; la première correspondait à la première année de suivi ; c'est pour cette période que des seuils de nadir de PSA ont été définis. Pour la seconde période, située au delà d'un an de suivi, il a été recommandé d'utiliser d'autres méthodes diagnostiques plus tardives, proposées par Blana et *al.* (Blana et *al.*, 2009). Néanmoins, en définissant d'autres valeurs de seuils de nadir, plus faibles, il serait peut être possible de détecter également les biopsies qui se positivent tardivement.

Le nadir a été considéré comme un marqueur mesuré de manière fixe, même si, d'un patient à l'autre, il n'est pas forcément atteint au même moment. Sa capacité à prédire dépend du délai écoulé entre le traitement et la date de la détection clinique de la persistance de

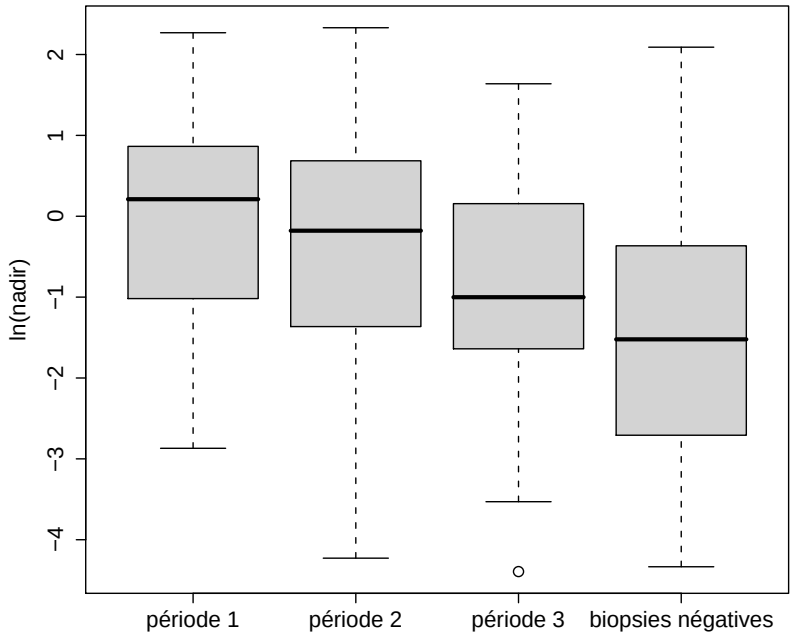

Figure 8.2 – Logarithme des valeurs de nadirs de PSA (ln(ng/mL)) en fonction de la période de positivation de la biopsie.

cellules cancéreuses. Pour les biomarqueurs mesurés de manière répétée au cours du temps et dont le niveau sert directement à effectuer le diagnostic précoce d'une maladie, la capacité du marqueur à prédire les événements peut dépendre de deux temps différents : le temps écoulé depuis l'inclusion dans l'étude et la mesure du biomarqueur, ainsi que le délai écoulé entre la mesure et la détection clinique de la maladie. Des mesures de sensibilité et de spécificité intégrant ces notions de temps sont donc nécessaires.

Dans le cadre des données de PSA, tous les patients ayant eu une biopsie positive au cours du suivi correspondent vraisemblablement à des patients pour lesquels il restait des cellules cancéreuses après le traitement (actives ou non). Ainsi, les valeurs de nadir de PSA sont utilisées pour effectuer le diagnostic précoce d'une maladie qui est présente dans un état latent, et non encore détectable cliniquement.

Plusieurs méthodes proposées pour tenir compte du temps dans l'estimation des performances diagnostiques sont présentées dans ce chapitre, en discutant des limites de chacune. Une ébauche de méthode d'estimation de seuils en fonction du temps est également décrite.

8.1 Performances diagnostiques dépendant du temps

La sensibilité mesure la capacité d'un marqueur à détecter les malades, la spécificité reflète quant à elle la capacité à ne pas classer comme malades des personnes non malades. Lorsque le statut des patients évolue au cours du temps, il est nécessaire de définir qui sont les " malades " et les " non malades " à un moment donné. Par la suite, on notera s le temps écoulé entre l'inclusion dans l'étude et la mesure du marqueur et t le délai entre la mesure et la détection clinique de la maladie pour les patients qui la présentent à un moment donné. Dans le cas du nadir de PSA, le nadir est supposé correspondre à une mesure fixe, effectuée à la même date pour tous les patients, car la variabilité des dates de nadir est relativement faible. Le temps écoulé s depuis le traitement est supposé ne pas avoir d'impact sur la capacité du nadir à prédire le résultat de la biopsie ; il est artificiellement fixé à zéro pour tous les patients.

Plusieurs méthodes ont été décrites par Heagerty et Zheng (2005) pour définir les malades et les non malades en fonction du temps, chacune étant associée à des mesures de sensibilité et de spécificité différentes. Elles ont été comparées récemment par Pepe et *al.* (2008) et Subtil et *al.* (2009).

8.1.1 Définition des malades

8.1.1.1 La sensibilité cumulative

Pour une mesure effectuée au temps s, les malades peuvent être définis comme étant les patients qui développent la maladie durant les t jours suivant la mesure du biomarqueur. La sensibilité dite *cumulative* est alors donnée par :

$$\mathrm{Sen}_{\mathbb{C}}(c, t, s) = P(Y(s) > c | T \leq t + s)$$

où T correspond au temps écoulé entre l'inclusion dans l'étude et la détection clinique de la maladie. La sensibilité cumulative est facile à estimer, car elle correspond simplement à la proportion de patients développant la maladie après la date $t+s$ et qui ont une valeur de marqueur au temps s supérieure au seuil de positivité considéré. Néanmoins, cette mesure de performance dépend fortement de la distribution des temps de survenue d'événements par rapport à l'inclusion dans l'étude.

Voici un exemple factice, où un marqueur est mesuré pour tous les patients au moment de l'inclusion, mais dont la valeur dépend de la période durant laquelle la maladie est détectée cliniquement :

- pour une maladie détectée durant les 10 premiers jours suivant l'inclusion, la valeur du marqueur à l'inclusion est de 15 ;
- pour une maladie détectée entre 10 et 20 jours, la valeur du marqueur est de 10 ;
- pour une maladie détectée entre 20 et 30 jours, la valeur du marqueur est de 5 ;
- enfin, le marqueur a une valeur nulle pour une maladie détectée au delà de 30 jours.

Le marqueur est donc d'autant plus élevé que la détection de la maladie est proche. Une étude est réalisée dans deux hôpitaux différents ; les résultats en sont présentés sur la figure 8.3. La flèche horizontale représente le temps écoulé depuis l'inclusion dans l'étude. Chaque caducée représente la date à laquelle un patient a développé la maladie, avec en dessous la valeur de marqueur au moment de l'inclusion.

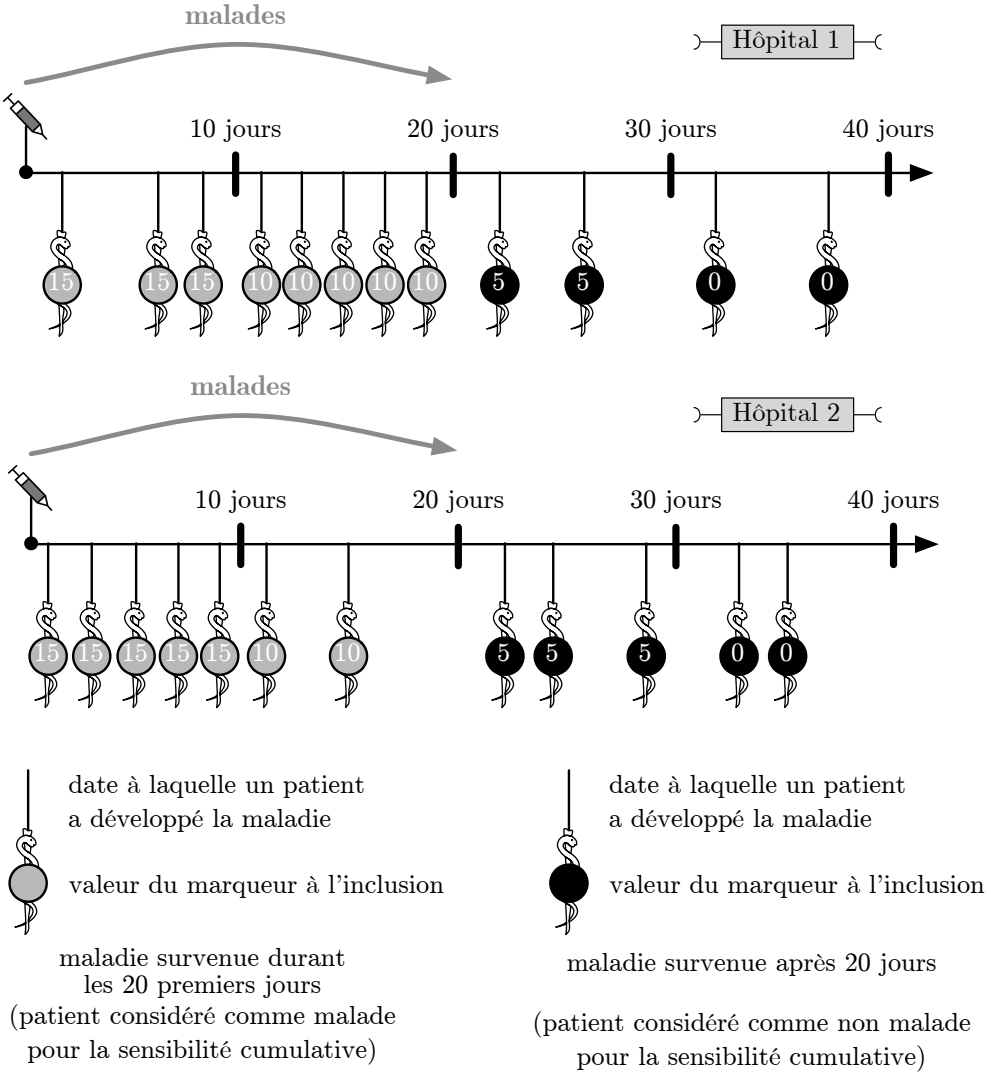

Figure 8.3 – Valeurs de marqueurs obtenues dans les deux établissements pour les patients considérés comme malades ou non malades à 20 jours de l'inclusion d'après la notion de sensibilité cumulative.

Supposons que l'on souhaite évaluer la capacité du marqueur à prédire la survenue de la maladie durant les 20 premiers jours suivant sa mesure. Les malades, selon la définition associée à la sensibilité cumulative, sont représentés en gris. La sensibilité du seuil de positivité 10 est :

- de 3/8 pour le premier établissement, car 3 patients ont une mesure de marqueur strictement supérieure à 10 parmi les 8 patients ayant développé la maladie durant les 20 premiers jours de suivi ;
- de 5/7 pour le second établissement, car 5 patients ont une mesure de marqueur strictement supérieure à 10 parmi les 7 patients ayant développé la maladie durant les 20 premiers jours de suivi.

Les valeurs de sensibilité cumulative estimées sont donc très différentes pour les deux établissements, alors que le même marqueur est considéré, pour un même délai d'analyse. La seule différence entre les deux hôpitaux est la distribution des temps de survenue d'événements. Pour le premier établissement, durant les 20 premiers jours, les patients qui développent la maladie la développent majoritairement tardivement ; ainsi, dans le groupe des malades, il y a plus de valeurs de marqueurs relativement faibles. A l'inverse, dans le second établissement, la majorité des patients qui développent la maladie durant les 20 premiers jours la développent relativement tôt ; les valeurs de marqueurs des malades sont donc en moyenne plus élevées. En cumulant les patients développant la maladie jusqu'à un certain temps, la sensibilité cumulative mélange des valeurs de marqueurs associées à des patients aux risques hétérogènes ; elle correspond à une moyenne de valeurs de marqueurs pour des stades différents d'avancement de la maladie. Un changement dans la distribution des temps de survenue d'événements entraîne un changement dans l'estimation de la sensibilité cumulative.

Ainsi, la sensibilité cumulative reflète :

– la capacité du marqueur à détecter les malades, ce qui est souhaité ;

– la distribution des dates d'événements au cours du temps.

Les estimations de sensibilités cumulatives obtenues d'une étude à l'autre risquent de ne pas être comparables, ce qui est problématique lorsque l'objectif est justement de comparer des marqueurs.

8.1.1.2 La sensibilité incidente

Pour la sensibilité incidente, les malades associés à une mesure effectuée au temps s correspondent aux patients qui développent la maladie exactement t jours après la mesure. Mathématiquement, la sensibilité incidente est donnée par :

$$\mathrm{Sen}_{\mathbb{I}}(c, t, s) = P(Y(s) > c | T = t + s)$$

Avec une telle définition, tous les patients considérés comme malades correspondent à des patients qui développent la maladie exactement au même temps après la mesure. Ce sont donc des patients au même niveau d'avancement de la maladie, dont les valeurs de marqueurs doivent être relativement homogènes (sauf si d'autres caractéristiques que le niveau d'avancement de la maladie agissent sur le niveau du marqueur). En utilisant la notion de sensibilité incidente, les per-

formances obtenues ne dépendent plus de la distribution des temps de survenue d'événements ; les résultats sont donc comparables d'une étude à l'autre.

L'inconvénient majeur de la sensibilité incidente est qu'elle n'est pas estimable par de simples proportions. En effet, la plupart des études n'incluent par suffisamment de patients pour que la proportion de personnes développant la maladie à un jour donné, et dont le marqueur est supérieur au seuil, soit suffisamment fiable. Il est donc nécessaire de faire appel à des techniques de modélisation. Ceci explique pourquoi la sensibilité cumulative est plus souvent utilisée, sans mentionner le fait que les résultats risquent de ne pas être comparables d'une étude à l'autre.

8.1.2 Définition des non malades

8.1.2.1 La spécificité dynamique

Pour la spécificité dynamique, les non malades pour une mesure effectuée aux temps s correspondent aux patients qui ne développent pas la maladie durant les t jours suivant la mesure. Mathématiquement, la spécificité dynamique est donnée par :

$$\mathrm{Spe}_{\mathbb{D}}(c, t, s) = P(Y(s) < c | T > t + s)$$

Un inconvénient de cette méthode est que le groupe des non malades varie en fonction du délai considéré, alors qu'il est préférable dans la plupart des études que ce groupe reste fixe au cours du temps.

8.1.2.2 La spécificité statique

Pour la spécificité statique, il est nécessaire de définir un temps $t^\star$ considéré comme suffisamment long pour que tous les patients qui ne développent pas la maladie avant $t^\star$ soient considérés comme des patients qui ne développeront jamais la maladie. La spécificité statique est donnée par :

$$\mathrm{Spe}_{\mathbb{S}}(c, t^\star, s) = P(Y(s) < c | T > t^\star + s)$$

De cette façon, la spécificité est calculée indépendamment des délais choisis pour l'estimation de la sensibilité.

8.1.3 Méthodes d'estimation

La sensibilité incidente ne peut pas être estimée à partir de simples proportions comme la sensibilité cumulative ; elle nécessite une phase de modélisation. De plus, la présence de patients perdus de vue, dont le statut n'est plus connu à partir d'un certain temps, peut poser problème dans les estimations de sensibilité et de spécificité.

Un certain nombre de méthodes ont été proposées dans la littérature pour tenir compte des censures lors de l'estimation de performances dépendant du temps, faisant appel très souvent aux méthodes utilisées dans le cadre de l'analyse de données de survie. Beaucoup de ces méthodes sont basées sur la définition de sensibilité cumulative, sans mentionner les inconvénients qui lui sont liés. Ces inconvénients n'ont été mis en évidence que tardivement, d'où l'importance de la discussion précédente.

Deux grands approches sont à distinguer pour l'estimation des performances dépendant du temps. La première repose principalement sur la modélisation des valeurs du marqueur en fonction de la date de la mesure et du délai de développement de la maladie (Zheng et Heagerty, 2004 ; Cai et *al.*, 2006). La seconde est basée sur le théorème de Bayes, permettant l'inversion des probabilités ; elle repose par conséquent sur la modélisation de la survie dans un délai donné suivant la mesure en fonction de la valeur du marqueur. Plusieurs méthodes ont été proposées :

- Zheng et *al.* (2006) utilisent le principe de pondération par l'inverse de la probabilité de non censure pour tenir compte de ces censures ;
- Zheng et Heagerty (2007) modélisent conjointement le statut du patient et l'évolution du marqueur ;
- la méthode proposée par Song et Zhou (2008) peut être adaptée à la fois à la sensibilité incidente et à la sensibilité statique ;
- celle de Cai et Cheng (2008) permet de combiner plusieurs marqueurs.

Toutes ces méthodes ont été développées par des équipes très proches ; les résultats fournis reposent très souvent sur des estimations obtenues à partir de la méthode des équations d'estimation généralisées (Liang et Zeger, 1986).

8.2 Seuil optimal du nadir de PSA en fonction du délai de positivation de la biopsie

8.2.1 Evaluation des performances diagnostiques du nadir de PSA en fonction du délai de positivation

Dans les chapitres précédents, les non malades ont été définis comme étant les patients sans biopsie positive durant la totalité du suivi. Ceci permet le calcul d'une spécificité statique. Pour la sensibilité, la notion de sensibilité incidente a été utilisée, mais pour simplifier les calculs, l'évaluation de cette sensibilité incidente a été réalisée selon les périodes définies au début de ce chapitre. La sensibilité dans la période 2 a été calculée à partir des valeurs de nadirs des patients dont la persistance de cancer à été détectée entre 170 et 480 jours. Ceci ne correspond pas à une estimation cumulative de sensibilité, qui aurait considéré l'ensemble des patients ayant eu une biopsie positive jusqu'à 480 jours. La méthode proposée d'estimation de la sensibilité mélange tout de même des patients aux risques hétérogènes, puisque la persistance de cellules cancéreuses est détectée à des dates différentes au cours d'une période. Néanmoins, avec le découpage en trois périodes, l'hétérogénéité entre les patients doit rester modérée.

Le tableau 8.1 présente les AROC pour les trois marqueurs considérés, à partir des estimations de sensibilité incidente et de spécificité statique obtenues sur les trois périodes d'études. La capacité à identifier les malades est d'autant plus élevée que la détection clinique de la maladie est précoce, d'où l'intérêt de proposer plusieurs seuils en fonction de la période considérée.

Tableau 8.1 – AROC (intervalles de confiance à 95 %) pour les différents marqueurs en fonction de la période d'évaluation des performances diagnostiques.

	Périodes confondues	Période 1	Période 2	Période 3
Nadir	0,692 [0,685 ; 0,699]	0,757 [0,747 ; 0,765]	0,710 [0,697 ; 0,721]	0,616 [0,605 ; 0,625]
Date du nadir	0,635 [0,596 ; 0,670]	0,707 [0,643 ; 0,759]	0,668 [0,611 ; 0,714]	0,539 [0,494 ; 0,593]
Vélocité	0,534 [0,486 ; 0,588]	0,583 [0,515 ; 0,652]	0,550 [0,492 ; 0,608]	0,473 [0,406 ; 0,534]

8.2.2 Estimation du seuil optimal de nadir de PSA en fonction du délai de positivation

Dans la partie sur l'estimation du seuil optimal du nadir de PSA, il a été considéré que la distribution du nadir suivait une loi log normale chez les malades et les non malades. Néanmoins, une régression linéaire du logarithme de la valeur du nadir en fonction de la date de la biopsie positive chez les malades montre un effet significatif de cette date sur les valeurs de nadir à chaque itération de l'algorithme MCMC, avec une diminution du nadir plus la biopsie positive est tardive. Ainsi, le logarithme du nadir chez les malades peut être modélisé grâce à la relation :

$$\ln(\text{nadir}_1) = \beta_0 + \beta_1 \times t_b \tag{8.1}$$

où t_b correspond à la date de la biopsie positive. En utilisant des a priori non informatifs pour β_0 et β_1, il est possible, grâce aux algorithmes de type MCMC, d'échantillonner dans la distribution a posteriori de ces paramètres, puis d'effectuer le calcul de la sensibilité pour un délai de positivation donné :

$$P(\text{nadir}_1 > c | T = t)$$

A partir de cette sensibilité, le seuil discriminant au mieux les non malades des patients dont la biopsie sera positive au jours t peut être calculé. Pour les patients dont le nadir est supérieur au seuil donné, les cliniciens savent qu'en effectuant une biopsie le jour t, il y a des chances non négligeables qu'elle soit positive. Il est néanmoins plus réaliste de définir un seuil au delà duquel la biopsie risque d'être positive durant une période donnée ($[\tau_1, \tau_2]$). Un tel seuil nécessite le calcul d'une sensibilité incidente sur une période donnée :

$$P(\text{nadir}_1 > c | \tau_1 \leq T < \tau_2) = \frac{\int_{\tau_1}^{\tau_2} P(\text{nadir}_1 > c | T = t) d(P(T = t))}{P(\tau_1 \leq T < \tau_2)}$$

Si t_i dénote la date de la biopsie positive du $i^{\text{ème}}$ patient, alors cette intégrale peut être approximée par :

$$P(\text{nadir}_1 > c | \tau_1 \leq T < \tau_2) \approx \frac{\sum_{i=1}^{n_1} P(\text{nadir}_1 > c | T = t_i) I(\tau_1 \leq t_i < \tau_2)}{\sum_{i=1}^{n_1} I(\tau_1 \leq t_i < \tau_2)}$$

où $P(\text{nadir}_1 > c | T = t_i)$ est obtenu à partir de l'équation (8.1) et des estimations de paramètres issues de la chaîne MCMC.

A partir de cette méthode, il serait possible d'obtenir des seuils pour les trois périodes de suivi présentées précédemment. Ainsi, pour un patient dont le nadir serait supérieur au seuil de la période un, le clinicien pourrait proposer une biopsie durant la première période. Un patient dont le nadir serait inférieur au seuil de la première période, mais supérieur à celui de la période deux, pourrait se voir proposer une biopsie durant la seconde période et ainsi de suite.

8.2.3 Prise en compte des censures

Dans l'analyse des données de PSA, les non malades correspondent aux patients dont toutes les biopsies sont négatives. Néanmoins, au delà de la dernière biopsie négative, il est difficile d'assurer de manière certaine que les patients sont toujours des non malades ; le suivi de ces patients devrait être censuré à la date de dernière biopsie.

Cai et *al.* (2006) ont proposé une méthode consistant à diviser le suivi des patients censurés en deux périodes pour l'estimation des performances diagnostiques d'un marqueur en présence de censures. Soit x_i la date de la dernière biopsie négative d'un patient considéré comme non malade et $t^\star$ la date au delà de laquelle un patient est considéré comme non malade de manière définitive. Après x_i, le patient peut continuer à être un non malade, et ce, avec une probabilité $P(T_i > t^\star | T_i > x_i)$. Il peut également développer la maladie à n'importe quel temps compris entre x_i et $t^\star$. Ainsi, la probabilité que la valeur du nadir soit supérieure à un seuil c sachant que le patient a été censuré au temps x_i est donnée par :

$$P(\text{nadir} > c | T_i > x_i) = (1 - \text{Spe}(c, t^\star))P(T_i > t^\star | T_i > x_i) + \int_{x_i}^{t^\star} \text{Sen}(c, t)\, dP(T_i \leq t | T_i > x_i) \tag{8.2}$$

Le patient participe donc à la fois à l'estimation de la spécificité et à l'estimation de la sensibilité pour tous les délais compris entre x_i et $t^\star$, mais avec une certaine probabilité dans les deux cas.

Cette méthode pourrait être adaptée pour le calcul des paramètres de distribution du nadir chez les malades et les non malades. Un patient censuré participe, dans ce cas, à l'estimation des paramètres de distribution du marqueur chez les non malades avec une probabilité $P(T_i > t^\star | T_i > x_i)$, ainsi qu'à celle des paramètres de distribution du marqueur chez les malades, avec une probabilité similaire à celle indiquée dans l'équation (8.2).

8.3 Perspectives

Cette partie a permis d'introduire la notion de temps dans l'estimation des performances diagnostiques du nadir de PSA et l'estimation d'un seuil de ce nadir. Dans tout le reste du travail de thèse, il a été considéré que le groupe des malades constitue un groupe homogène, alors qu'il y a, dans les faits, une réelle hétérogénéité liée à la date à laquelle la persistance de cellules cancéreuses est détectée.

Ainsi, il peut être utile d'introduire non pas un unique seuil de nadir de PSA, mais plusieurs en fonction de la période durant laquelle une biopsie sera conseillée. Une ébauche de méthode a été proposée afin d'estimer ces seuils, en tenant compte du phénomène de censure, mais un certain nombre de difficultés techniques persistent, notamment pour le calcul des paramètres de distribution du marqueur en présence de censures. Une perspective est de pouvoir proposer, par la suite, plusieurs seuils de nadir de PSA en fonction de la période d'intérêt, ceci permettant d'optimiser les dates de réalisation de biopsies, et de limiter le nombre important de biopsies négatives.

Cette partie a également été l'occasion de présenter les différentes méthodes d'évaluation des performances diagnostiques dépendant du temps, et surtout, d'insister sur les limites de l'approche " sensibilité cumulative ", limites très peu mentionnées dans la littérature alors que la sensibilité cumulative est fréquemment utilisée.

Chapitre

9

Le coût du risque

Dans la partie concernant l'évaluation des performances d'un biomarqueur, il a été signalé que le risque de maladie associé aux différents niveaux du marqueur est fréquemment analysé, car il s'interprète facilement par le patient ou le clinicien. Connaissant la probabilité de maladie prédite par le marqueur, un médecin prend une décision concernant le traitement éventuel d'un patient, en intégrant dans son raisonnement les caractéristiques spécifiques de ce patient, par exemple en termes d'état de santé général ou d'aversion vis à vis de la maladie et des traitements. Pour que la décision soit la plus juste possible, il est important que le risque prédit par le marqueur corresponde bien au risque réel du patient. Néanmoins, le processus d'intégration des bénéfices et des coûts du traitement par le médecin reste relativement flou. Des méthodes rationnelles simples sont nécessaires pour, d'une part, aider le patient à formuler ses préférences en termes de qualité ou de quantité de vie, et d'autre part, intégrer correctement ces préférences dans le processus de prise de décision. C'est ce second aspect qui est abordé ici.

Lorsqu'un marqueur est retenu selon l'analyse des courbes de prédiction, il est conseillé de choisir celui qui assigne au plus de personnes possibles des risques élevés ou faibles et non intermédiaires. Mais cette démarche n'intègre pas les coûts et les bénéfices d'un traitement suite à la mesure du marqueur.

Ces deux remarques posent la question de l'intégration de l'utilité d'une action dans la modélisation du risque associé aux différents niveaux d'un biomarqueur. Une amorce de réponse est proposée dans ce chapitre.

9.1 Le risque de maladie pour l'estimation du seuil optimal d'un marqueur

La modélisation du risque d'une maladie en fonction d'une ou plusieurs covariables est très fréquente en recherche clinique et en épidémiologie. De nombreux outils ont été développés dans ce sens, dont le plus connu est certainement la régression logistique. Une méthode d'estimation du seuil optimal reposant sur le risque de maladie sachant la valeur du marqueur serait par conséquent facilement utilisable à grande échelle.

DeBari (2006) a proposé en ce sens de retenir la valeur du marqueur telle que la probabilité de maladie associée soit de 1/2, sans justification concrète. Néanmoins, cette proposition, sous réserve de quelques modifications, peut avoir du sens. Soit Y la valeur d'un marqueur. D'après le théorème de Bayes, la probabilité de maladie pour une valeur Y peut s'écrire :

$$P(M=1|Y) = \frac{P(Y|M=1)P(M=1)}{P(Y|M=1)P(M=1) + P(Y|M=0)P(M=0)} \tag{9.1}$$

Ainsi, la valeur de Y correspondant au seuil proposé par DeBari est telle que :

$$P(M=1|Y) = \frac{P(Y|M=1)P(M=1)}{P(Y|M=1)P(M=1) + P(Y|M=0)P(M=0)} = \frac{1}{2} \tag{9.2}$$

ce qui équivaut à :

$$\frac{P(Y|M=1)}{P(Y|M=0)} = \frac{1-P(M=1)}{P(M=1)}$$

Le seuil proposé correspond à la valeur du marqueur telle que le ratio de la densité de probabilité chez les malades et les non malades en ce point soit égal à l'inverse de la cote de maladie. Dans le chapitre sur l'estimation du seuil, il a été montré que le seuil maximisant l'utilité espérée correspond à la valeur du marqueur telle que (équation 5.1) :

$$\frac{P(Y|M=1)}{P(Y|M=0)} = \frac{CN}{BN} \times \frac{(1-P(M=1))}{P(M=1)} \tag{9.3}$$

Ainsi, le seuil proposé par DeBari correspond au seuil maximisant l'utilité espérée lorsque le ratio bénéfice sur coût net est de 1.

Cette méthode peut être adaptée pour d'autres valeurs de ratio bénéfice net sur coût net, afin de tenir compte des préférences individuelles des patients. Rechercher la valeur de seuil qui

vérifie la relation (9.3) revient à déterminer la valeur Y telle que :

$$\frac{P(Y|M=1)P(M=1)\times BN}{P(Y|M=1)P(M=1)\times BN + P(Y|M=0)P(M=0)\times CN} = \frac{1}{2} \tag{9.4}$$

Cette équation est très similaire à l'équation (9.2), sauf que les utilités BN et CN sont introduites au numérateur et au dénominateur. Ainsi, les malades ont un poids de BN, correspondant au bénéfice net à être traités à raison ; les non malades ont un poids de CN, correspondant au coût net à être traités à tort. Au lieu de rechercher la valeur de Y telle que le risque de maladie soit de $1/2$, on recherche la valeur de Y tel que le risque pondéré de maladie soit de $1/2$.

Les programmes dédiés à la réalisation de régressions logistiques contiennent très souvent une option permettant d'introduire des poids, ce qui revient à augmenter ou diminuer artificiellement le nombre de malades ou de non malades associés à un niveau du marqueur. Dans ce cas, il ne faut interpréter que les estimations ponctuelles des coefficients et non leur erreur type, car les poids n'introduisent pas d'information supplémentaire en termes de nombre de patients, mais simplement des bénéfices ou des coûts liés à l'action de traiter une personne.

Au lieu d'effectuer une régression logistique pondérée, il est également possible d'effectuer une régression logistique simple, mais en cherchant cette fois ci la valeur du marqueur telle que :

$$P(M=1|Y) = R^\star \quad \text{avec} \quad R^\star = \frac{CN}{CN+BN} \tag{9.5}$$

Les équations (9.2) et (9.5) sont équivalentes.

9.2 Risque de maladie et chances de guérir

La régression logistique pondérée présentée en (9.4) est équivalente, en termes d'estimations ponctuelles, à une régression logistique introduisant en offset la valeur $\ln(BN/CN)$. Ainsi, la probabilité pondérée de maladie est équivalente en termes de cotes et de rapport de cotes à :

$$\text{Cote}_Y = \frac{BN}{CN} \times \text{Cote}_0 \times \text{RC}^Y \tag{9.6}$$

où RC est le rapport de cotes associé au marqueur Y et Cote_0 correspond à la cote pour une valeur nulle du marqueur. Cette méthode a l'avantage de ne pas modifier les erreurs types des estimations ponctuelles.

L'équation (9.6) montre que le fait d'introduire des coûts dans la modélisation du risque de maladie ne modifie pas le niveau d'association entre le marqueur et la maladie, puisque le rapport de cotes reste inchangé ; l'introduction des utilités modifie uniquement la cote de base. Ainsi, les conclusions quant aux choix de marqueurs ou de combinaisons de marqueurs effectuées uniquement en termes de rapport de cotes ne changent pas en introduisant l'utilité de l'action.

Un exemple est proposé dans le cas où le marqueur suit une loi normale de moyenne -1 chez les non malades, 1 chez les malades et d'écart type 1 dans les deux groupes. La figure 9.1 représente les probabilités pondérées de maladie en fonction de la mesure du marqueur, pour trois valeurs différentes du ratio bénéfice net sur coût net.

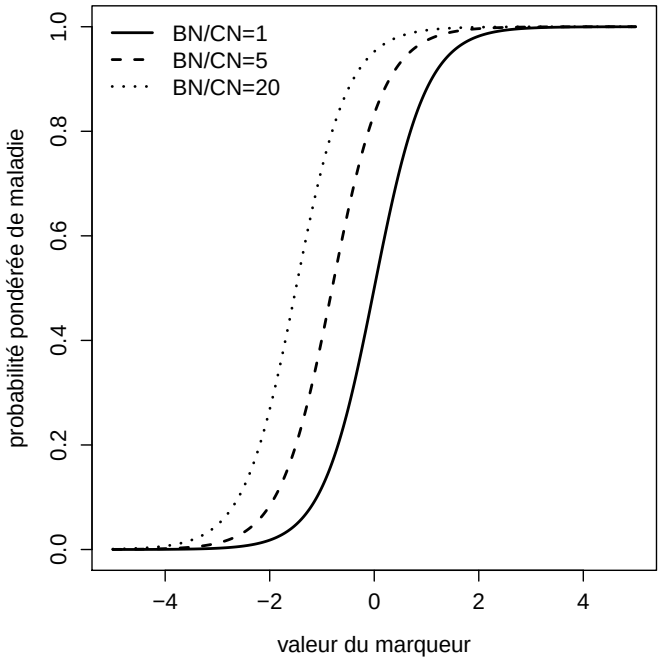

Figure 9.1 – Probabilité pondérée de maladie en fonction de la valeur du marqueur pour trois valeurs différentes de ratio BN/CN.

Le fait de modifier les utilités ne modifie pas la capacité du marqueur à discriminer les malades des non malades, les courbes gardant la même inclinaison, mais modifie le niveau à partir duquel le marqueur considère que l'individu est " malade ". Plus le ratio bénéfice net sur coût net augmente, plus le bénéfice à être traité à raison est grand par rapport au coût d'être traité à tort ; ainsi, la probabilité de maladie prédite se rapproche rapidement de un, même pour des valeurs faibles de marqueur. La probabilité pondérée, tenant compte du bénéfice à être traité à raison et du coût à être traité à tort, peut s'interpréter comme la probabilité que le

traitement soit bénéfique à partir d'une valeur du marqueur. Ainsi, plus le bénéfice à être traité est élevé, plus le marqueur a intérêt à ce qu'une grande partie de la population soit traitée ; par conséquent, la probabilité que l'action de traiter soit positive passe plus rapidement à un.

La probabilité que l'action de traiter soit positive peut être plus simple à utiliser qu'une probabilité de maladie. En effet, une fois le ratio bénéfice net sur coût net déterminé pour un patient (ou la valeur de risque justifiant un traitement ou une intervention), la probabilité que le traitement ait une action bénéfique intègre directement ces coûts et bénéfices et pourrait être lue sur des abaques selon la valeur du marqueur.

Cette méthode serait également utile dans les études pronostiques, où fréquemment, plusieurs seuils du marqueur – ou combinaisons de niveaux de marqueurs – sont déterminés en fonction de niveaux de risques. Ici encore, l'utilisation de la probabilité que l'action d'une intervention soit positive au lieu de la simple probabilité de maladie pourrait rendre plus facile la définition des différents seuils de risques critiques, ainsi que, par conséquent, celle des différents seuils du marqueur.

9.3 Perspectives

Cette partie n'est qu'une ébauche de réflexion sur l'introduction des utilités dans l'estimation du risque. Néanmoins, les pistes évoquées semblent prometteuses ; de nombreux problèmes pourraient bénéficier de cette approche. L'interprétation des mesures d'adéquation de modèles de risque lorsque celui-ci intègre l'utilité des actions n'a pas encore été abordée.

Concernant l'estimation du seuil optimal d'un marqueur, il faut constater que, bien souvent, le seuil est défini de manière empirique, ou uniquement en maximisant l'indice de Youden, qui ne tient compte ni de la prévalence de la maladie, ni du ratio bénéfice net sur coût net. Les méthodes les plus simples sont les méthodes les plus utilisées, même si elles ne sont pas toujours les plus adaptées. La méthode d'estimation du seuil basée sur la modélisation du risque de maladie associé aux différents niveaux du marqueur est une méthode maximisant l'utilité espérée, mais qui reste simple à mettre en œuvre, le monde de la recherche clinique et épidémiologique étant habitué à la modélisation du risque. Avec une telle méthode, les seuils de décision retenus pour les nouveaux biomarqueurs pourraient devenir des seuils réellement “ utiles ” pour les patients.

Conclusion

Dans ce travail de thèse, un ensemble de méthodes a été proposé concernant l'utilisation des biomarqueurs quantitatifs longitudinaux pour le diagnostic précoce ou le pronostic. Ces méthodes vont du choix du critère – issu de la cinétique du marqueur – qui est le plus approprié pour discriminer les patients, à la détermination du seuil pour proposer un traitement. Pour toutes ces méthodes, l'accent a été mis sur trois aspects :

- obtenir les estimations les plus justes possibles des différents critères par patient afin que le choix du critère et le choix du seuil ne soient pas influencés, par exemple, par des problèmes de fréquence ou d'erreurs de mesures ;
- tenir compte de toutes les sources d'incertitude et de variabilité dans les estimations ponctuelles et les intervalles de crédibilité des paramètres ;
- choisir un critère et un seuil qui ne minimisent pas le nombre moyen d'erreurs de classement, mais à partir desquels les décisions prises conduiront en moyenne à la plus grande utilité, en termes d'état de santé, pour la population considérée.

L'inférence Bayésienne a été utilisée tout au long de ce travail. Pour la partie concernant la modélisation des profils de biomarqueurs, ce choix était surtout lié à la simplicité des algorithmes MCMC dans le cas de modèles complexes. Concernant l'estimation du seuil optimal, l'inférence Bayésienne est très adaptée pour fournir un intervalle de crédibilité qui tienne compte de toutes les formes d'incertitude.

Toutes les démarches précédentes ont été appliquées dans le cas de l'utilisation des mesures de PSA pour déclencher des biopsies après traitement UFHI, pour des patients présentant une suspicion de persistance de cellules cancéreuses après trois mois de suivi, avec le choix du critère issu de la cinétique des PSA le plus discriminant (le nadir de PSA) ainsi que la détermination de seuils du nadir de PSA. Ces seuils de nadir correspondent à des seuils de décision, valables en moyenne pour une population et non optimaux pour un individu. Le choix parmi les différentes valeurs de seuils dépend néanmoins des spécificités du patient pour lequel une décision doit être prise, spécificités en termes de caractéristiques cliniques ou biologiques, ou spécificités en

termes d'aversion vis-à-vis de la maladie et de ses traitements. Si ce travail de thèse propose des méthodes pour tenir compte de la variabilité de ses préférences, il n'aborde pas, ou peu, les méthodes utilisées pour amener les patients à les exprimer. Ceci constitue un pan entier de l'aide à la décision médicale. Une réflexion supplémentaire serait nécessaire afin d'établir des scénarios pour aider les cliniciens à définir, avec le patient, le risque de maladie justifiant la réalisation de la biopsie. Cette réflexion est indispensable pour que les seuils définis dans l'article destiné à la revue The European Urology soient réellement utilisés en pratique.

La méthode Bayésienne d'estimation du seuil optimal est valable à la fois pour des marqueurs issus de la cinétique d'un biomarqueur longitudinal, ou pour des marqueurs fixes. Elle peut donc avoir de nombreuses applications. Toute la difficulté sera de la communiquer dans le monde de la recherche clinique et de la clinique. En effet, bien souvent, les seuils de biomarqueurs retenus sont arbitraires, ou uniquement basés sur les quantiles des distributions dans la population ou sur l'indice de Youden. Ce travail permettra peut être de faire évoluer les pratiques dans ce domaine, afin que les décisions prises pour les patients par rapport aux valeurs de marqueurs correspondent bien à des décisions utiles.

Liste des références

Andriole G. L., Crawford E. D., Grubb R. L., Buys S. S., Chia D., Church T. R., Fouad M. N., Gelmann E. P., Kvale P. A., Reding D. J., Weissfeld J. L., Yokochi L. A., O'Brien B., Clapp J. D., Rathmell J. M., Riley T. L., Hayes R. B., et al. (2009). Mortality results from a randomized prostate-cancer screening trial. *The New England Journal of Medicine* **360,** 1310–1319.

ASTRO (1997). Consensus statement: guidelines for PSA following radiation therapy. American Society for Therapeutic Radiology and Oncology Consensus Panel. *International journal of radiation oncology, biology, physics* **37,** 1035–1041.

Attouch H. (1984). *Variational convergence for functions and operators.* Longman Higher Education.

Bauvin E., Remontet L., Grosclaude P., réseau FRANCIM, et CépiDc (2003). Incidence and mortality of prostate cancer in France: trends between 1978 and 2000. *Progrès en Urologie* **13,** 1334–1339.

Bellera C. A., Hanley J. A., Joseph L., et Albertsen P. C. (2008). Hierarchical changepoint models for biochemical markers illustrated by tracking postradiotherapy prostate-specific antigen series in men with prostate cancer. *Annals of Epidemiology* **18,** 270–282.

Berger J. O. (1980). *Statistical decision theory and Bayesian analysis.* Springer, 2nd edition.

Berkson J. et Gage R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47,** 501–515.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology Therapy* **69,** 89–95.

Black M. A. et Craig B. A. (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* **21,** 2653–2669.

Blana A., Brown S. C. W., Chaussy C., Conti G. N., Eastham J. A., Ganzer R., Murat F. J., Pasticier G., Rebillard X., Rewcastle J. C., Robertson C. N., Thuroff S., et Ward J. F. (2009). High-intensity focused ultrasound for prostate cancer: comparative definitions of biochemical failure. *BJU International* **104,** 1058–1062.

Blana A., Murat F. J., Walter B., Thuroff S., Wieland W. F., Chaussy C., et Gelet A. (2008a). First analysis of the long-term results with transrectal HIFU in patients with localised prostate cancer. *European Urology* **53,** 1194–1201.

Blana A., Rogenhofer S., Ganzer R., Lunz J.-C., Schostak M., Wieland W. F., et Walter B. (2008b). Eight years' experience with high-intensity focused ultrasonography for treatment of localized prostate cancer. *Urology* **72,** 1329–1333.

Bradlaw E., Lenk P., Alleby G., et Rossi P. (2004). *Market research and modeling: progress and prospects: a tribute to Paul E. Green* Marketing research – Methodology, pages 7–42. Springer, New York, 1st edition.

Brown E. R. et Ibrahim J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59,** 221–228.

Bush C. et MacEachern S. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83,** 275–286.

Cai T. et Cheng S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9,** 216–233.

Cai T., Pepe M. S., Zheng Y., Lumley T., et Jenny N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7,** 182–197.

Cao D.-L. et Yao X.-D. (2010). Advances in biomarkers for the early diagnosis of prostate cancer. *Chinese Journal of Cancer* **29,** 229–233.

Carroll R. J. et Ruppert D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

Casella G. et Berger R. (2002). *Statistical Inference*. Duxbury Resource Center, Belmont, 2nd edition.

Chaloner K. et Verdinelli I. (1995). Bayesian experimental design: a review. *Statistical Science* **10,** 273–304.

Chen J., Zhang D., et Davidian M. (2002). A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics* **3,** 347–360.

Choi Y.-K., Johnson W. O., Collins M. T., et Gardner I. A. (2006). Bayesian inferences for Receiver Operating Characteristic Curves in the absence of a gold standard. *Journal of Agricultural, Biological and Environmental Statistics* **11,** 210–229.

Ciezki J. P. et Klein E. A. (2009). Brachytherapy or surgery ? A composite view. *Oncology (Williston Park)* **23,** 960–964.

Clemen R. T. et Winkler R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis* **19,** 187–203.

Coffin M. et Sukhatme S. (1997). Receiver operating characteristic studies and measurement errors. *Biometrics* **53,** 823–837.

Condorcet N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris.

Cook N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115,** 928–935.

Critz F. A., Williams W. H., Holladay C. T., Levinson A. K., Benton J. B., Holladay D. A., Schnell F. J., Maxa L. S., et Shrake P. D. (1999). Post-treatment PSA < or = 0.2 ng/mL defines disease freedom after radiotherapy for prostate cancer using modern techniques. *Urology* **54,** 968–971.

Crook J. M., Perry G. A., Robertson S., et Esche B. A. (1995). Routine prostate biopsies following radiotherapy for prostate cancer : results for 226 patients. *Urology* **45,** 624–631.

Daskivich T. J., Regan M. M., et Oh W. K. (2006). Prostate specific antigen doubling time calculation: not as easy as 1, 2, 4. *The Journal of Urology* **176,** 1927–1937.

DeBari V. A. (2006). Computation of decision levels from differentiated logistic regression probability curves. *Annals of clinical and laboratory science* **36,** 194–200.

DeNeef P. et Kent D. L. (1993). Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis. *Medical Decision Making* **13,** 126–132.

Diamond G. A. (1992). What price perfection ? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology* **45,** 85–89.

Dorfman J. H. (1997). *Bayesian economis through numerical methods: a guide to econometrics and decision-making with prior information.* Springer, New-York.

Eastham J. A., Riedel E., Scardino P. T., Shike M., Fleisher M., Schatzkin A., Lanza E., Latkany L., et Begg C. B. (2003). Variation of serum prostate-specific antigen levels: an evaluation of year-to-year fluctuations. *The Journal of the American Medical Association* **289,** 2695–2700.

Efron B. et Tibshirani R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560.

Escobar M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89,** 268–277.

Escobar M. D. et West M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Fisher R. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society* **22,** 700–725.

Fluss R., Faraggi D., et Reiser B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal* **47,** 458–472.

Garthwaite P. H., Kadane J. B., et O'Hagan A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* **100,** 680–700.

Gelman A., Carlin J. B., Stern H. S., et Rubin D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, London, 2nd edition.

Geman S. et Geman D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6,** 721–741.

Gerds T. A., Cai T., et Schumacher M. (2008). The performance of risk prediction models. *Biometrical Journal* **50,** 457–479.

Gerszten R. E. et Wang T. J. (2008). The search for new cardiovascular biomarkers. *Nature* **451,** 949–952.

Geyer C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **56,** 261–274.

Gilks W., Richardson S., et Spiegelhalter D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Grosclaude P., Bossard N., Remontet L., Belot A., Arveux P., Bouvier A., Launoy G., Maynaié M., Velten M., Faivre J., et Estève J. (2007). *Survie des patients atteints de cancer en France - Etude des registres du réseau FRANCIM*. Springer-Verlag, Paris.

Gumbel E. (1966). *Statistics of extremes*. Columbia university press, New-York.

Hand D. J. (1997). *Construction and assessment of classification rules*. John Wiley & Sons, Chichester.

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

Hastings W. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57,** 97–109.

Heagerty P. J. et Zheng Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61,** 92–105.

Hoeting J. A., Madigan D., Raftery A. E., et Volinsky C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14,** 382–417.

Hosmer D. et Lemeshow S. (1989). *Applied logistic regression.* John Wiley & Sons, New York.

Huang Y., Pepe M. S., et Ziding F. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63,** 1181–1188.

Hull G. W., Rabbani F., Abbas F., Wheeler T. M., Kattan M. W., et Scardino P. T. (2002). Cancer control with radical prostatectomy alone in 1000 consecutive patients. *Journal of Urology* **167,** 528–534.

Ibrahim J. G. et Chen M.-H. (2000). Power prior distributions for regression models. *Statistical Science* **15,** 46–60.

Jeffreys H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences.* **186,** 453–461.

Joseph L., Gyorkos T. W., et Coupal L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* **141,** 263–272.

Jouini M. N. et Clemen R. T. (1996). Copula models for aggregating expert opinions. *Operations Research* **44,** 444–457.

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

Kall P. (1986). Approximation to optimization problems: an elementary review. *Mathematics of Operations Research* **11,** 9–18.

Kleinman K. P. et Ibrahim J. G. (1998). A semiparametric Bayesian approach to the random effects model. *Biometrics* **54,** 921–938.

Kohlmann M., Held L., et Grunert V. P. (2009). Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal* **51,** 610–626.

Laird N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Laird N. M. et Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lange K. L., Little R. J. A., et Taylor J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84,** 881–896.

Lavine M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics* **20,** 1222–1235.

Lavine M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics* **22,** 1161–1176.

Lawless J. F. (1996). *Statistical models and methods for lifetime data.* Wiley, Hoboken, 1st edition.

Lee K. J. et Thompson S. G. (2008). Flexible parametric models for random-effects distributions. *Statistics in Medicine* **27,** 418–434.

Lee S.-H., Chen S.-M., Ho C.-R., Chang P.-L., Chen C.-L., et Tsui K.-H. (2009). Risk factors associated with transrectal ultrasound guided prostate needle biopsy in patients with prostate cancer. *Chang Gung Medical Journal* **32,** 623–627.

Leeflang M. M. G., Moons K. G. M., Reitsma J. B., et Zwinderman A. H. (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical chemistry* **54,** 729–737.

Liang K. et Zeger S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin D. Y., Wei L. J., et Ying Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58,** 1–12.

Lipsitz S. R., Ibrahim J., et Molenberghs G. (2000). Using a Box-Cox transformation in the analysis of longitudinal data with incomplete responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49,** 287–296.

Lowrance W. T., Elkin E. B., Jacks L. M., Yee D. S., Jang T. L., Laudone V. P., Guillonneau B. D., Scardino P. T., et Eastham J. A. (2010). Comparative effectiveness of prostate cancer surgical treatments : a population based analysis of postoperative outcomes. *The Journal of Urology* **183,** 1366–1372.

MacEachern S. N. et Müller P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7,** 223–238.

McInturff P., Johnson W. O., Cowling D., et Gardner I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in Medicine* **23,** 1095–1109.

Metropolis N., Rosenbluth A., Rosenbluth M., Teller A., et Teller E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21,** 1087–1092.

Mittlböck M. et Schemper M. (1996). Explained variation for logistic regression. *Statistics in Medicine* **15,** 1987–1997.

Moons K. G. M. et Harrell F. E. (2003). Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology* **10,** 670–672.

Morris P. (1974). Decision analysis expert use. *Management Science* **20,** 1233–1241.

Morris P. (1977). Combining expert judgements: a Bayesian approach. *Management Science* **23,** 679–693.

Moskowitz C. S. et Pepe M. S. (2004). Quantifying and comparing the accuracy of binary biomarkers when predicting a failure time outcome. *Statistics in Medicine* **23,** 1555–1570.

Muller P. (1999). Simulation based optimal design. In Bernardo J., Berger J., Dawid A., et Smith A., editors, *Bayesian Statistics 6: Proceedings of the sixth Valencia international meeting*, pages 459–474, New-York. Oxford University Press.

Muller P. et Parmigiani G. (1996). *Bayesian analysis in statistics and econometrics: essays in honor of Arnold Zellner* Numerical evaluation of information theoretic measures, pages 397–406. Wiley, New York.

Muller P. et Quintana F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science* **19,** 95–110.

Murat F.-J., Gelet A., Belot A., et Rabilloud M. (2010). Residual prostate cancer after a first HIFU session. Oncologic benefits of a second session. (*article en cours de soumission*).

Murat F.-J., Poissonnier L., Rabilloud M., Belot A., Bouvier R., Rouviere O., Chapelon J.-Y., et Gelet A. (2009). Mid-term results demonstrate salvage high-intensity

focused ultrasound (HIFU) as an effective and acceptably morbid salvage treatment option for locally radiorecurrent prostate cancer. *European Urology* **55,** 640–647.

Murphy A. H. et Winkler R. L. (1987). A general framework for forecast verification. *Monthly weather review* **115,** 1330–1338.

Neal R. M. (2000). Markov chain sampling methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

Neyman J. et Pearson E. (1933). On the problem of the most efficient test of statistical hypotheses. *Philosophical Transactions of the Royal Society* **231,** 286–237.

O'Hagan A., Buck C. E., Daneshkhah A., Eiser J. R., Garthwaite P. H., Jenkinson D. J., Oakley J. E., et Rakow T. (2006). *Unceraint judgements: eliciting experts' probabilities.* Wiley, Chichester.

Ohlssen D. I., Sharples L. D., et Spiegelhalter D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* **26,** 2088–2112.

Parmigiani G. (2002). *Modeling in medical decision making. A Bayesian approach.* Wiley, Chichester.

Pasticier G., Chapet O., Badet L., Ardiet J. M., Poissonnier L., Murat F. J., Martin X., et Gelet A. (2008). Salvage radiotherapy after high-intensity focused ultrasound for localized prostate cancer: early clinical results. *Urology* **72,** 1305–1309.

Pencina M. J., Agostino R. B. D., Agostino R. B. D., et Vasan R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27,** 157–172.

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

Pepe M. S., Feng Z., Huang Y., Longton G., Prentice R., Thompson I. M., et Zheng Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology* **167,** 362–368.

Pepe M. S., Zheng Y., Jin Y., Huang Y., Parikh C. R., et Levy W. C. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis* **14,** 86–113.

Pinheiro J. C., Liu C., et Wu Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* **10,** 249–276.

Poissonnier L., Chapelon J.-Y., Rouvière O., Curiel L., Bouvier R., Martin X., Dubernard J. M., et Gelet A. (2007a). Control of prostate cancer by transrectal HIFU in 227 patients. *European Urology* **51,** 381–387.

Proust-Lima C. et Taylor J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* .

Pulkstenis E. et Robinson T. J. (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics in Medicine* **21,** 79–93.

Ransohoff D. F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer* **4,** 309–314.

Ray M. E., Levy L. B., Horwitz E. M., Kupelian P. A., Martinez A. A., Michalski J. M., Pisansky T. M., Zelefsky M. J., Zietman A. L., et Kuban D. A. (2006a). Nadir prostate-specific antigen within 12 months after radiotherapy predicts biochemical and distant failure. *Urology* **68,** 1257–1262.

Ray M. E., Thames H. D., Levy L. B., Horwitz E. M., Kupelian P. A., Martinez A. A., Michalski J. M., Pisansky T. M., Shipley W. U., Zelefsky M. J., Zietman A. L., et Kuban D. A. (2006b). PSA nadir predicts biochemical and distant failures after external beam radiotherapy for prostate cancer: a multi-institutional analysis. *International Journal of Radiation oncology, biology, physics* **64,** 1140–1150.

Roach M., Hanks G., Thames H., Schellhammer P., Shipley W. U., Sokol G. H., et Sandler H. (2006). Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. *International journal of radiation oncology, biology, physics* **65,** 965–974.

Robert C. P. et Casella G. (1999). *Monte Carlo statistical methods.* Springer, New-York.

Rosa G., Padovani C., et Gianola D. (2003). Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. *Biometrical Journal* **45,** 573–590.

Rouvière O., Girouin N., Glas L., Cheikh A. B., Gelet A., Mège-Lechevallier F., Rabilloud M., Chapelon J.-Y., et Lyonnet D. (2010). Prostate cancer transrectal HIFU ablation: detection of local recurrences using T2-weighted and dynamic contrast-enhanced MRI. *European Radiology* **20,** 48–55.

Rustin G. J. S. (2003). Use of CA-125 to assess response to new agents in ovarian cancer trials. *Journal of Clinical Oncology* **21,** 187s–193s.

Savage L. (1954). *The Foundations of Statistics*. Wiley, New York.

Scarpa B. et Dunson D. B. (2007). Bayesian methods for searching for optimal rules for timing intercourse to achieve pregnancy. *Statistics in Medicine* **26,** 1920–1936.

Schiffer E. (2009). The 2nd annual oncology biomarkers conference. *Biomarkers in Medicine* **3,** 203–209.

Schisterman E. F., Faraggi D., Reiser B., et Hu J. (2008). Youden Index and the optimal threshold for markers with mass at zero. *Statistics in Medicine* **27,** 297–315.

Schisterman E. F. et Perkins N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics: Simulation and Computation* **36,** 549–563.

Schröder F. H., Hugosson J., Roobol M. J., Tammela T. L. J., Ciatto S., Nelen V., Kwiatkowski M., Lujan M., Lilja H., Zappa M., Denis L. J., Recker F., et al. (2009). Screening and prostate-cancer mortality in a randomized European study. *The New England journal of medicine* **360,** 1320–1328.

Slate E. H. et Turnbull B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine* **19,** 617–637.

Soletormos G., Semjonow A., Sibley P. E., Lamerz R., Petersen P. H., Albrecht W., Bialk P., Gion M., Junker F., Schmid H.-P., Van Poppel H., et on behalf of the

European Group on Tumor Markers (2005). Biological variation of total prostate-specific antigen: a survey of published estimates and consequences for clinical practice. *Clinical chemistry* **51,** 1342–1351.

Song X. et Zhou X.-H. (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinca* **18,** 947–966.

Subtil F., Pouteil-Noble C., Toussaint S., Villar E., et Rabilloud M. (2009). A simple modeling-free method provides accurate estimates of sensitivity and specificity of longitudinal disease biomarkers. *Methods of Information in Medicine* **48,** 299–305.

Taylor J. M. G., Yu M., et Sandler H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23,** 816–825.

Tsiatis A. A., DeGruttola V., et Wulfsohn M. S. (1995). Modeling the relationship of survival to longitudinal data measures with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90,** 27–37.

Tu X. M., Kowalski J., et Jia G. (1999). Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening. *Statistics in Medicine* **18,** 3059–3073.

Uchida T., Shoji S., Nakano M., Hongo S., Nitta M., Murota A., et Nagata Y. (2009). Transrectal high-intensity focused ultrasound for the treatment of localized prostate cancer: eight-year experience. *International Journal of Urology* **16,** 881–886.

Verbeke G. et Lesaffre E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis* **23,** 541–556.

Vickers A. J., Cronin A. M., Elkin E. B., et Gonen M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Information Decision Making* **8,** 53.

Vickers A. J. et Elkin E. B. (2006). Decision curve analysis, a novel method for evaluating prediction models. *Medical Decision Making* **26,** 565–574.

Wang M.-D. et Geisser S. (2005). Optimal dichotomization of screening test variables. *Journal of statistical planning and inference* **131,** 191–206.

Winkler R. (1981). Combining probability distributions from dependent information sources. *Management Science* **27,** 479–488.

You B., Perrin P., Freyer G., Ruffion A., Tranchand B., Hénin E., Paparel P., Ribba B., Devonec M., Falandry C., Fournel C., Tod M., et Girard P. (2008). Advantages of prostate-specific antigen (PSA) clearance model over simple PSA half-life computation to describe PSA decrease after prostate adenomectomy. *Clinical Biochemistry* **41,** 785–795.

Zelefsky M. J., Shi W., Yamada Y., Kollmeier M. A., Cox B., Park J., et Seshan V. E. (2009). Postradiotherapy 2-year prostate-specific antigen nadir as a predictor of long-term prostate cancer mortality. *International journal of radiation oncology, biology, physics* **75,** 1350–1356.

Zheng Y., Cai T., et Feng Z. (2006). Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62,** 279–287.

Zheng Y. et Heagerty P. J. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5,** 615–632.

Zheng Y. et Heagerty P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics* **63,** 332–341.

Zolg J. W. et Langen H. (2004). How industry is approaching the search for new diagnostic test markers and biomarkers ? *Molecular and Cellular Proteomics* **3,** 345–354.

Annexe

A

Inférence Bayésienne

A.1 Principes

Soit $\mathbf{y}$ des données pouvant être décrites grâce à un modèle de paramètres $\boldsymbol{\theta}$. L'objectif de l'inférence Bayésienne est de modéliser la distribution des paramètres sachant les données. D'après le théorème de Bayes :

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})}{P(\mathbf{y})} = \frac{P(\mathbf{y}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})}{\int_{\Theta} P(\mathbf{y}|\boldsymbol{\theta}) \times P(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

$P(\boldsymbol{\theta}|\mathbf{y})$ est la probabilité a posteriori de $\boldsymbol{\theta}$, sachant les données. $P(\mathbf{y}|\boldsymbol{\theta})$ correspond à la vraisemblance des observations et $P(\boldsymbol{\theta})$ est la probabilité a priori des paramètres, c'est à dire la probabilité supposée des valeurs de $\boldsymbol{\theta}$ avant que les mesures ne soient réalisées. Cette probabilité peut être spécifiée à partir d'analyses antérieures sur le même sujet. Lorsque les données contiennent beaucoup d'information, le poids de l'information a priori est négligeable par rapport à la vraisemblance dans la distribution a posteriori ; il est alors possible d'utiliser un a priori non informatif, consistant à donner a priori la même probabilité à toutes les valeurs de $\boldsymbol{\theta}$.

Un des avantages majeurs de l'inférence Bayésienne est qu'elle ne fournit pas uniquement une estimation ponctuelle des paramètres, mais l'intégralité de leur distribution a posteriori. Une estimation ponctuelle peut être obtenue en calculant le mode, la médiane ou la moyenne de la distribution a posteriori. Un intervalle de crédibilité à 95 % peut être construit en retenant les quantiles 2,5 et 97,5 % de la distribution a posteriori (méthode appelée dans ce travail méthode des quantiles), ou bien en déterminant la région de plus haute densité de probabilité (méthode

appelée dans ce travail méthode HDP). Cette région contient 95 % de la densité a posteriori, la densité de probabilité des points lui appartenant étant au moins aussi élevée que celle des points ne lui appartenant pas. Par rapport à la méthode des quantiles, l'intervalle de crédibilité obtenu par la méthode HDP peut être disjoint, par exemple dans le cas de distributions bimodales.

La définition d'un intervalle de crédibilité est différente de celle d'un intervalle de confiance. L'intervalle de crédibilité au niveau α est un intervalle tel que la probabilité que la vraie valeur du paramètre lui appartienne sachant les données est de $1-\alpha$. L'intervalle de confiance fréquentiste est relié au hasard d'échantillonnage : si l'échantillonnage des observations était ré-effectué un grand nombre de fois, sous l'hypothèse que le mécanisme de génération des données ne change pas, alors 95 % des valeurs estimées du paramètre seraient comprises dans l'intervalle de confiance fréquentiste. Lorsque le nombre de mesures est élevé, intervalle de crédibilité et intervalle de confiance coïncident.

Il existe des a priori conjugués à la vraisemblance des données, de telle sorte que la distribution a posteriori est connue de manière explicite ; par exemple, la loi beta est conjuguée à la vraisemblance binomiale. Ainsi, si l'information a priori pour la probabilité d'un événement est spécifiée sous la forme d'une loi beta, alors la distribution a posteriori de la probabilité de l'événement suit également une loi beta. La loi gamma est conjuguée quant à elle à la vraisemblance de Poisson.

Dans d'autres cas, la densité a posteriori est calculable en tout point, mais la loi a posteriori n'est pas connue de manière explicite. Une solution consiste alors à échantillonner un grand nombre de valeurs selon la distribution de probabilité a posteriori, puis à calculer les statistiques d'intérêt à partir de l'échantillon de valeurs ainsi obtenu. C'est le principe de l'intégration de Monte Carlo. Toute la difficulté est de pouvoir échantillonner dans une distribution de probabilité lorsque celle-ci n'est pas connue de manière explicite. Les algorithmes de type Monte Carlo - chaîne de Markov (MCMC pour Monte Carlo Markov Chain) sont couramment utilisés dans ce type de situation.

Une chaîne de Markov est une succession de variables aléatoires $(\theta^1, \ldots, \theta^i, \ldots, \theta^n)$ telle que, pour tout i, la distribution de θ^i sachant les valeurs qui la précèdent ne dépend que de la valeur la plus récente (θ^{i-1}). Les règles de transition permettant le passage d'une itération à l'autre d'une chaîne MCMC sont choisies de telle sorte que, au bout d'un certain temps, la distribution des valeurs de la chaîne converge vers une distribution stationnaire, correspondant à la distribution a posteriori du ou des paramètres d'intérêt. La clé du succès de la méthode n'est pas la propriété de Markov, mais le fait qu'à chaque itération, le tirage est ajusté de

sorte que la distribution se rapproche de plus en plus de la distribution d'intérêt. Deux des principaux algorithmes MCMC sont présentés par la suite. Ces algorithmes ne sont pour autant pas spécifiques de l'inférence Bayésienne.

A.2 L'échantillonneur de Gibbs

L'échantillonneur de Gibbs est l'une des méthodes MCMC fréquemment employées en inférence Bayésienne (Geman et Geman, 1984). Il est utilisé pour échantillonner dans la distribution a posteriori de plusieurs paramètres. Le principe consiste, à chaque itération, à échantillonner dans les distributions de probabilité conditionnelles de chacun des paramètres. Dans le cas où $\boldsymbol{\theta}$ est de dimension k, l'algorithme est le suivant :

1. prendre des valeurs de départ pour $\boldsymbol{\theta} = (\theta_1^0, \ldots, \theta_k^0)^T$;
2. pour $i = 1, 2, \ldots$
   - tirer θ_1^i selon $P(\theta_1 | \theta_2^{i-1}, \ldots, \theta_k^{i-1}, \mathbf{y})$;
   - tirer θ_2^i selon $P(\theta_2 | \theta_1^i, \theta_3^{i-1}, \ldots, \theta_k^{i-1}, \mathbf{y})$;
   - tirer θ_k^i selon $P(\theta_k | \theta_1^i, \ldots, \theta_{k-1}^i, \mathbf{y})$.

L'algorithme est répété jusqu'à ce que la chaîne converge vers une distribution stationnaire. Il est souvent plus aisé d'échantillonner dans les distributions conditionnelles que dans la distribution a posteriori simultanée de l'ensemble des paramètres.

A.3 Le Metropolis-Hastings

Le Metropolis-Hastings est un algorithme permettant d'échantillonner dans la distribution a posteriori d'un ou de plusieurs paramètres (Metropolis et *al.*, 1953 ; Hastings, 1970). Le principe consiste à explorer de manière aléatoire l'espace des valeurs possibles des paramètres et à retenir préférentiellement les valeurs associées à une densité de probabilité a posteriori élevée. Le Metropolis-Hastings nécessite une distribution de proposition $P(\boldsymbol{\theta}^\star | \boldsymbol{\theta}^{i-1})$, permettant de générer une valeur candidate de $\boldsymbol{\theta}$ à partir de la valeur retenue à l'itération précédente. Cette distribution de proposition est arbitraire ; une loi normale multivariée centrée sur la valeur de $\boldsymbol{\theta}$ à l'itération précédente est couramment utilisée. La valeur candidate $\boldsymbol{\theta}^\star$ est retenue avec une probabilité $\alpha^\star$ définie par :

$$\alpha^\star = \min\left(\frac{P(\boldsymbol{\theta}^\star|\mathbf{y})P(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^\star)}{P(\boldsymbol{\theta}^{i-1}|\mathbf{y})P(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^{i-1})}, 1\right)$$

Si $\boldsymbol{\theta}^\star$ est rejetée, la valeur de $\boldsymbol{\theta}$ à l'itération $i-1$ est conservée pour l'itération i.

L'échantillonneur de Gibbs et le Metropolis-Hastings peuvent être combinés dans un même algorithme, en échantillonnant par exemple des valeurs dans les distributions conditionnelles grâce au Metropolis-Hastings.

A.4 Intégration de Monte Carlo et théorème central limite

Soit Y une variable aléatoire de densité de probabilité $f(y)$ et de fonction de répartition $F(y)$, et $g(y)$ une fonction à valeurs réelles dans $\mathbb{R}$. Soit $(y_1, \ldots, y_n)$ une série de réalisations *indépendantes* de y. Alors l'intégrale :

$$I = \int_0^1 g(y)\, dF(y)$$

peut être approximée par :

$$E = \frac{1}{n} \sum_{i=1}^{n} g(y_i)$$

Par la loi des grands nombres, $\lim_{n\to\infty} \sum_{i=1}^{n} g(y_i)/n = I$. Si les réalisations de y sont indépendantes, alors d'après le théorème central limite, $\sqrt{n}(E-I)$ tend, en distribution, vers une loi normale de variance :

$$\int_0^1 (g(y) - I)^2\, dF(y)$$

Un estimateur de cette variance est donné par :

$$\sum_{i=1}^{n} ((g(y_i))^2 - E^2)/n$$

Ces résultats ne sont plus valables lorsque les réalisations de la variable aléatoire ne sont pas indépendantes. C'est par exemple le cas lorsque cette variable est un paramètre dont des valeurs ont été échantillonnées dans la distribution a posteriori via une méthode MCMC. En effet, les tirages issus d'un algorithme MCMC ne sont pas indépendants les uns des autres.

Dans la partie portant sur l'estimation du seuil optimal du marqueur (partie 5.2.1.2), une méthode est présentée permettant d'utiliser, sous certaines conditions, le théorème central limite à partir d'échantillons obtenus via une méthode MCMC.

Annexe

B

Annexe concernant le chapitre 4

Supplementary materials: Robust non-linear mixed model for longitudinal PSA measurements after a treatment

Fabien Subtil[1] and Muriel Rabilloud[1]

May 2009

[1] Université de Lyon, F-69000, Lyon, France.
Université Lyon 1, F-69001, Lyon, France.
CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique - Santé, F-69622, Villeurbanne, France.
Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France.

Correspondence to: Fabien Subtil (`fabien.subtil@chu-lyon.fr`)

1 Sampling from the posterior for the Gauss | Gauss model

We used Gibbs sampling to sample from the joint posterior distribution of the parameters : $r1$, $r2$, $r3$, $r4$, μ_{r_1}, μ_{r_2}, μ_{r_3}, μ_{r_4}, $\sigma^2_{r_1}$, $\sigma^2_{r_2}$, $\sigma^2_{r_3}$, $\sigma^2_{r_4}$ and σ^2_{ε}. The joint posterior doesn't have closed form ; however, given that the conditionnal posteriors either have closed form or can be sampled using a Metropolis, the implementation using the Gibbs sampler was straightforward. Let D denotes the data, *rest* denotes the remaining parameters, and N the total number of measurements over patients.

The following notations were used for the hyperpriors:

$$\mu_{r_1} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_1}}, \sigma^2_{\mu_{r_1}}) \quad \mu_{r_2} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_2}}, \sigma^2_{\mu_{r_2}}) \quad \mu_{r_3} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_3}}, \sigma^2_{\mu_{r_3}}) \quad \mu_{r_4} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_4}}, \sigma^2_{\mu_{r_4}})$$

$$1/\sigma^2_{r_1} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_1}}, \beta_{\sigma_{r_1}}) \quad 1/\sigma^2_{r_2} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_2}}, \beta_{\sigma_{r_2}})$$

$$1/\sigma^2_{r_3} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_3}}, \beta_{\sigma_{r_3}}) \quad 1/\sigma^2_{r_4} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_4}}, \beta_{\sigma_{r_4}})$$

$$1/\sigma^2_{\varepsilon} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{\varepsilon}}, \beta_{\sigma_{\varepsilon}})$$

Let :

$$\psi(t_{ij}) = \ln\left(\exp(r_{1i})\exp(-r_{2i}t_{ij}) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1\right)$$

Then, at each iteration of the Gibbs sampler, we proceed as follows:

1. Sample $[1/\sigma_{\varepsilon}^2|rest, D]$ from

$$\text{Gamma}\left(\frac{\alpha_{\sigma_\varepsilon}+N}{2}, \frac{\beta_{\sigma_\varepsilon}+\sum_{i=1}^{n}\sum_{j=1}^{m_j}(\ln(y_{ij}+1)-\psi(t_{ij}))^2}{2}\right)$$

2. Sample $[r_{1i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{1i} at the previous iteration, and with variance $\sigma^2_{metro_{r1}}$. The log posterior $[r_{1i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma_{\varepsilon}^2}\sum_{j=1}^{m_i}\left(\psi^2(t_{ij})-2y_{ij}\psi(t_{ij})\right)+\frac{r_{1i}^2-2r_{1i}\mu_{r_1}}{\sigma_{r_1}^2}\right]$$

3. Sample $[r_{2i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{2i} at the previous iteration, and with variance $\sigma^2_{metro_{r2}}$. The log posterior $[r_{2i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma_{\varepsilon}^2}\sum_{j=1}^{m_i}\left(\psi^2(t_{ij})-2y_{ij}\psi(t_{ij})\right)+\frac{r_{2i}^2-2r_{2i}\mu_{r_2}}{\sigma_{r_2}^2}\right]$$

4. Sample $[r_{3i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{3i} at the previous iteration, and with variance $\sigma^2_{metro_{r3}}$. The log posterior $[r_{3i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma_{\varepsilon}^2}\sum_{j=1}^{m_i}\left(\psi^2(t_{ij})-2y_{ij}\psi(t_{ij})\right)+\frac{r_{3i}^2-2r_{3i}\mu_{r_3}}{\sigma_{r_3}^2}\right]$$

5. Sample $[r_{4i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{3i} at the previous iteration, and with variance $\sigma^2_{metro_{r4}}$. The posterior $[r_{4i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma_{\varepsilon}^2}\sum_{j=1}^{m_i}\left(\psi^2(t_{ij})-2y_{ij}\psi(t_{ij})\right)+\frac{r_{4i}^2-2r_{4i}\mu_{r_4}}{\sigma_{r_4}^2}\right]$$

6. Sample $[\mu_{r_1}|rest, D]$ from $\mathcal{N}\left(\left(\sum_{i=1}^{n} r_{1i}+\mu_{\mu_{r_1}}\right)\sigma^2_{\mu_{r_1}}/(\sigma^2_{\mu_{r_1}}N+\sigma^2_{r_1}), \sigma^2_{\mu_{r_1}}\sigma^2_{r_1}/(\sigma^2_{\mu_{r_1}}N+\sigma^2_{r_1})\right)$.

7. Sample $[\mu_{r_2}|rest, D]$ from $\mathcal{N}\left(\left(\sum_{i=1}^{n} r_{2i}+\mu_{\mu_{r_2}}\right)\sigma^2_{\mu_{r_2}}/(\sigma^2_{\mu_{r_2}}N+\sigma^2_{r_2}), \sigma^2_{\mu_{r_2}}\sigma^2_{r_2}/(\sigma^2_{\mu_{r_2}}N+\sigma^2_{r_2})\right)$.

8. Sample $[\mu_{r_3}|rest, D]$ from $\mathcal{N}\left(\left(\sum_{i=1}^{n} r_{3i}+\mu_{\mu_{r_3}}\right)\sigma^2_{\mu_{r_3}}/(\sigma^2_{\mu_{r_3}}N+\sigma^2_{r_3}), \sigma^2_{\mu_{r_3}}\sigma^2_{r_3}/(\sigma^2_{\mu_{r_3}}N+\sigma^2_{r_3})\right)$.

9. Sample $[\mu_{r_4}|rest, D]$ from $\mathcal{N}\left(\left(\sum_{i=1}^{n} r_{4i}+\mu_{\mu_{r_4}}\right)\sigma^2_{\mu_{r_4}}/(\sigma^2_{\mu_{r_4}}N+\sigma^2_{r_4}), \sigma^2_{\mu_{r_4}}\sigma^2_{r_4}/(\sigma^2_{\mu_{r_4}}N+\sigma^2_{r_4})\right)$.

10. Sample $[1/\sigma^2_{r_1}|rest, D]$ from Gamma $\left((n+\alpha_{\sigma_{r_1}})/2, (\beta_{\sigma_{r_1}}+\sum_{i=1}^{n}(r_{1i}-\mu_{r_1})^2)/2\right)$.

11. Sample $[1/\sigma^2_{r_2}|rest, D]$ from Gamma $\left((n+\alpha_{\sigma_{r_2}})/2, (\beta_{\sigma_{r_2}}+\sum_{i=1}^{n}(r_{2i}-\mu_{r_2})^2)/2\right)$.

12. Sample $[1/\sigma^2_{r_3}|rest, D]$ from Gamma $\left((n+\alpha_{\sigma_{r_3}})/2, (\beta_{\sigma_{r_3}}+\sum_{i=1}^{n}(r_{3i}-\mu_{r_3})^2)/2\right)$.

13. Sample $[1/\sigma^2_{r_4}|rest, D]$ from Gamma $\left((n+\alpha_{\sigma_{r_4}})/2, (\beta_{\sigma_{r_4}}+\sum_{i=1}^{n}(r_{4i}-\mu_{r_4})^2)/2\right)$.

2 Sampling from the posterior for the Student | Dirichlet model

Using the equivalence between the Student-t distribution and a mixture of normal ones, the Student | Dirichlet model can be rewritten as follow :

$$\ln(y_{ij}+1) = \ln(\exp(r_{1i})\exp(-r_{2i}t_{ij})) + \exp(r_{3i})\exp(r_{4i}t_{ij}) + 1) + \varepsilon_{ij}$$

$$\varepsilon_{ij} \hookrightarrow \mathcal{N}(0, V_{ij}) \qquad V_{ij} \hookrightarrow \text{Inv} - \chi^2(\nu, \sigma^2)$$

$$r_{1i} \hookrightarrow DP(M_1 G_{01}) \quad r_{2i} \hookrightarrow \mathcal{N}(\mu_{r_2}, \sigma^2_{r_2}) \quad r_{3i} \hookrightarrow \mathcal{N}(\mu_{r_3}, \sigma^2_{r_3}) \quad r_{1i} \hookrightarrow DP(M_4 G_{04})$$

$$G_{01} \hookrightarrow \mathcal{N}(\mu_{r_1}, \sigma^2_{r_1}) \quad G_{04} \hookrightarrow \mathcal{N}(\mu_{r_4}, \sigma^2_{r_4})$$

2.1 Sampling from a Dirichlet process using Neal 8 algorithm

(a careful reading of Neal's article is recommended for this part).

Sampling from a Dirichlet process (for example for r_1) consists in clustering r_1 values in a set of k clusters $\phi_1, \ldots, \phi_k$; every r_1 belonging to the same cluster have the same value ϕ. The number of clusters can change from one iteration of the Gibbs sampler to another, and r_1 values can also change from one cluster to another during the Gibbs sampler. Let k be the number of clusters at one iteration, and $c_1, \ldots, c_n$ the cluster to which belongs each r_1 value (each patient) at the same iteration. Hence, the algorithm consists in sampling ϕ_j depending on the likelihood of data belonging to the j^{th} cluster, and then to draw a new value for c_i depending on the $\phi_1, \ldots, \phi_k$ values that have been updated. In reality, at each iteration, an r_1 value can also belong to a cluster that didn't exist at the previous iteration, whose ϕ value is sampled from G_0. Neal 8 algorithm consists in sampling at each iteration m auxiliary ϕ values from G_0, and then drawing a new value for c_i from $\{1, \ldots, k, k+1, k+m\}$ using the following probabilities:

$$P(c_i = c | c_{-i}, D, \phi_1, \ldots, \phi_{k+m}, rest) = \begin{cases} b\frac{n_{-i,c}}{n-1+M} F(y_i, \phi_c) & \text{for } 1 \leq c \leq k \\ b\frac{M/m}{n-1+M} F(y_i, \phi_c) & \text{for } k < c \leq k+m \end{cases}$$

where c_{-i} denote the set of cluster number for each r_1 value except the i^{th} one; $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c, b is the appropriate normalizing constant, and $F(y_i, \phi_c)$ denotes the likelihood of the measurements of the i^{th} patient under the hypothesis that r_{1i} takes the value ϕ_c.

2.2 Gibbs sampler

We used Gibbs sampling to sample from the joint posterior distribution of the parameters : ϕ_1, $r2$, $r3$, ϕ_4, μ_{r_1}, μ_{r_2}, μ_{r_3}, μ_{r_4}, V, $\sigma^2_{r_1}$, $\sigma^2_{r_2}$, $\sigma^2_{r_3}$, $\sigma^2_{r_4}$, σ^2, M_1, M_4 and ν. Let D denotes the data, $rest$ denotes the remaining parameters, N the total number of measurements over patients, and k_1 and k_4 the number of clusters for r_1 and r_4 (numbers that change from one iteration to another).

The following notations were used for the hyperpriors:

$$\mu_{r_1} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_1}}, \sigma^2_{\mu_{r_1}}) \quad \mu_{r_2} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_2}}, \sigma^2_{\mu_{r_2}}) \quad \mu_{r_3} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_3}}, \sigma^2_{\mu_{r_3}}) \quad \mu_{r_4} \hookrightarrow \mathcal{N}(\mu_{\mu_{r_4}}, \sigma^2_{\mu_{r_4}})$$

$$1/\sigma^2_{r_1} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_1}}, \beta_{\sigma_{r_1}}) \quad 1/\sigma^2_{r_2} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_2}}, \beta_{\sigma_{r_2}})$$

$$1/\sigma^2_{r_3} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_3}}, \beta_{\sigma_{r_3}}) \quad 1/\sigma^2_{r_4} \hookrightarrow \text{Gamma}(\alpha_{\sigma_{r_4}}, \beta_{\sigma_{r_4}})$$

$$M_1 \hookrightarrow \text{Gamma}(\alpha_{M_1}, \beta_{M_1}) \quad M_4 \hookrightarrow \text{Gamma}(\alpha_{M_4}, \beta_{M_4})$$

1. Sample λ from $\chi^2_{\nu+1}$ and let $V_{ij} = \lambda / \left(\nu\sigma^2 + (\ln(y_{ij}+1) - \psi(t_{ij}))^2\right)$

2. Sample $[\phi_{1i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of ϕ_{1i} at the previous iteration, and with variance $\sigma^2_{metro_{\phi_1}}$. The log posterior $[\phi_{1i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\sum_{l\in\Omega_{\phi_{1i}}}\sum_{j=1}^{m_l}\frac{\psi^2_{\phi_{1i}}(t_{lj}) - 2y_{lj}\psi_{\phi_{1i}}(t_{lj})}{V_{lj}} + \frac{\phi^2_{1i} - 2\phi_{1i}\mu_{r_1}}{\sigma^2_{r_1}}\right]$$

   where $\Omega_{\phi_{1i}}$ denotes the set of patients whom r_1 value belongs to cluster ϕ_{1i} at the previous iteration, and

$$\psi_{\phi_{1i}}(t_{lj}) = \ln\left(\exp(\phi_{1i})\exp(-r_{2l}t_{lj}) + \exp(r_{3l})\exp(r_{4l}t_{lj}) + 1\right)$$

3. Sample $[c_i|c_{-i}, D, rest]$ in two steps:

   (a) Sample m values of ϕ_1 in G_{01} ($\mathcal{N}(\mu_{r_1}, \sigma^2_{r_1})$), denoted $\phi_{1(k_1+1)}, \ldots, \phi_{1(k_1+m)}$.

   (b) Draw a new value for c_i in $\{1, \ldots, k_1, k_1+1, \ldots, k_1+m\}$ using the following probabilities:

$$P(c_i = c|c_{-i}, D, \phi_{11}, \ldots, \phi_{1(k_1+m)}, rest) = \begin{cases} b\frac{n_{-i,c}}{n-1+M_1}F(y_i, \phi_{1c}) & \text{for } 1 \le c \le k_1 \\ b\frac{M_1/m}{n-1+M_1}F(y_i, \phi_{1c}) & \text{for } k_4 < c \le k_1+m \end{cases}$$

   with $F(y_i, \phi_{1c}) = \prod_{j=1}^{m_i} 1/\sqrt{(}2\pi V_i j) \times \exp\left(-0.5 \times (\ln(y_{ij}+1) - \psi_{\phi_{1c}}(t_{ij}))^2/V_{ij}\right)$, and b the appropriate normalizing constant.

4. Let $r_{1i} = \phi_{1c_i}$.

5. Sample $[r_{2i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{2i} at the previous iteration, and with variance $\sigma^2_{metro_{r2}}$. The log posterior $[r_{2i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\sum_{j=1}^{m_i}\frac{\left(\psi^2(t_{ij}) - 2y_{ij}\psi(t_{ij})\right)}{V_{ij}} + \frac{r^2_{2i} - 2r_{2i}\mu_{r_2}}{\sigma^2_{r_2}}\right]$$

6. Sample $[r_{3i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of r_{3i} at the previous iteration, and with variance $\sigma^2_{metro_{r3}}$. The log posterior $[r_{3i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\sum_{j=1}^{m_i}\frac{\left(\psi^2(t_{ij}) - 2y_{ij}\psi(t_{ij})\right)}{V_{ij}} + \frac{r_{3i}^2 - 2r_{3i}\mu_{r_3}}{\sigma^2_{r_3}}\right]$$

7. Sample $[\phi_{4i}|rest, D]$ using a Metropolis with normal proposal, centered around the value of ϕ_{4i} at the previous iteration, and with variance $\sigma^2_{metro_{\phi_4}}$. The log posterior $[\phi_{4i}|rest, D]$ is given by:

$$-\frac{1}{2}\left[\sum_{l\in\Omega_{\phi_{4i}}}\sum_{j=1}^{m_l}\frac{\psi^2_{\phi_{4i}}(t_{lj}) - 2y_{lj}\psi_{\phi_{4i}}(t_{lj})}{V_{lj}} + \frac{\phi_{4i}^2 - 2\phi_{4i}\mu_{r_4}}{\sigma^2_{r_4}}\right]$$

where $\Omega_{\phi_{4i}}$ denotes the set of patients whom r_4 value belongs to cluster ϕ_{4i} at the previous iteration.

8. Sample $[c_i|c_{-i}, D, rest]$ in two steps:

   (a) Sample m values of ϕ_4 in G_{04} ($\mathcal{N}(\mu_{r_4}, \sigma^2_{r_4})$), denoted $\phi_{4(k_4+1)}, \ldots, \phi_{4(k_4+m)}$.

   (b) Draw a new value for c_i in $\{1, \ldots, k_4, k_4+1, \ldots, k_4+m\}$ using the following probabilities:

$$P(c_i = c|c_{-i}, D, \phi_{41}, \ldots, \phi_{4(k_4+m)}, rest) = \begin{cases} b\frac{n_{-i,c}}{n-1+M_4}F(y_i, \phi_{4c}) & \text{for } 1 \le c \le k_4 \\ b\frac{M_4/m}{n-1+M_4}F(y_i, \phi_{4c}) & \text{for } k_4 < c \le k_4+m \end{cases}$$

   with $F(y_i, \phi_{4c}) = \prod_{j=1}^{m_i}\left(\ln(y_{ij}+1) - \psi_{\phi_{4c}}(t_{ij})\right)^2/V_{ij}$, and b the appropriate normalizing constant.

9. Let $r_{4i} = \phi_{4c_i}$.

10. Sample M_1 in two steps:

    (a) Sample the latent variable η from Beta(M_1+1, k_1) (where Beta(a, b) is the Beta distribution).

    (b) Sample M_1 from:

$$[M_1|\eta, k_1] \hookrightarrow \pi_\eta \text{Gamma}(\alpha_{M_1} + k_1) + (1-\pi_\eta)\text{Gamma}(\beta_{M_1} - \ln(\eta))$$

    where $\pi_\eta/(1-\pi_\eta) = (\alpha_{M_1} + k_1 - 1)/(n(\beta_{M_1} - \ln(\eta)))$.

11. Sample M_4 in two steps:

    (a) Sample the latent variable η from Beta(M_4+1, k_4).

    (b) Sample M_4 from:

$$[M_4|\eta, k_4] \hookrightarrow \pi_\eta \text{Gamma}(\alpha_{M_4} + k_4) + (1-\pi_\eta)\text{Gamma}(\beta_{M_4} - \ln(\eta))$$

    where $\pi_\eta/(1-\pi_\eta) = (\alpha_{M_4} + k_4 - 1)/(n(\beta_{M_4} - \ln(\eta)))$.

12. Sample $[\mu_{r_1}|rest, D]$ from $\mathcal{N}\left(\left(\sigma^2_{\mu_{r_1}}\sum_{i=1}^{k_1}\phi_{1i} + \sigma^2_{r_1}\mu_{\mu_{r_1}}\right)/(\sigma^2_{\mu_{r_1}}k_1 + \sigma^2_{r_1}), (\sigma^2_{r_1}\sigma^2_{\mu_{r_1}})/(\sigma^2_{\mu_{r_1}}k_1 + \sigma^2_{r_1})\right)$.
13. Sample $[\mu_{r_2}|rest, D]$ from $\mathcal{N}\left(\left(\sigma^2_{\mu_{r_2}}\sum_{i=1}^{n}r_{2i} + \sigma^2_{r_2}\mu_{\mu_{r_2}}\right)/(\sigma^2_{\mu_{r_2}}n + \sigma^2_{r_2}), (\sigma^2_{r_2}\sigma^2_{\mu_{r_2}})/(\sigma^2_{\mu_{r_2}}n + \sigma^2_{r_2})\right)$.
14. Sample $[\mu_{r_3}|rest, D]$ from $\mathcal{N}\left(\left(\sigma^2_{\mu_{r_3}}\sum_{i=1}^{n}r_{3i} + \sigma^2_{r_3}\mu_{\mu_{r_3}}\right)/(\sigma^2_{\mu_{r_3}}n + \sigma^2_{r_3}), (\sigma^2_{r_3}\sigma^2_{\mu_{r_3}})/(\sigma^2_{\mu_{r_3}}n + \sigma^2_{r_3})\right)$.
15. Sample $[\mu_{r_4}|rest, D]$ from $\mathcal{N}\left(\left(\sigma^2_{\mu_{r_4}}\sum_{i=1}^{k_4}\phi_{4i} + \sigma^2_{r_4}\mu_{\mu_{r_4}}\right)/(\sigma^2_{\mu_{r_4}}k_4 + \sigma^2_{r_4}), (\sigma^2_{r_4}\sigma^2_{\mu_{r_4}})/(\sigma^2_{\mu_{r_4}}k_4 + \sigma^2_{r_4})\right)$.
16. Sample $[1/\sigma^2_{r_1}|rest, D]$ from Gamma $\left((k_1 + \alpha_{r_1})/2, \left(\sum_{i=1}^{k_1}(\phi_{1i} - \mu_{r_1})^2 + \beta_{r_1}\right)/2\right)$.
17. Sample $[1/\sigma^2_{r_2}|rest, D]$ from Gamma $\left((n + \alpha_{r_2})/2, \left(\sum_{i=1}^{n}(r_{2i} - \mu_{r_2})^2 + \beta_{r_2}\right)/2\right)$.
18. Sample $[1/\sigma^2_{r_3}|rest, D]$ from Gamma $\left((n + \alpha_{r_3})/2, \left(\sum_{i=1}^{n}(r_{3i} - \mu_{r_3})^2 + \beta_{r_3}\right)/2\right)$.
19. Sample $[1/\sigma^2_{r_4}|rest, D]$ from Gamma $\left((k_4 + \alpha_{r_4})/2, \left(\sum_{i=1}^{k_4}(\phi_{4i} - \mu_{r_4})^2 + \beta_{r_4}\right)/2\right)$.
20. Sample $[1/\sigma^2|rest, D]$ from Gamma $\left((N\nu)/2, \nu/2 \times \sum_{i=1}^{n}\sum_{j=1}^{m_i}1/V_{ij}\right)$.
21. Sample $[\nu|rest, D]$ using a Metropolis with normal proposal, centered around the value of ν at the previous iteration, and with variance $\sigma^2_{metro_\nu}$. The log posterior $[\nu|rest, D]$ is given by:

$$N\left(\frac{\nu}{2}\ln\left(\frac{\nu}{2}\right) + \nu\ln(\sigma) - \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right)\right) - \left(\frac{\nu}{2} + 1\right)\sum_{i=1}^{n}\sum_{j=1}^{m_i}\ln(V_{ij}) - \frac{\nu\sigma^2}{2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}V_{ij}$$

We chose $\sigma_{metro_{r1}} = 0.1$, $\sigma_{metro_{r2}} = 0.001$ $\sigma_{metro_{r3}} = 0.7$ $\sigma_{metro_{r4}} = 0.002$ and $\sigma_{metrop_\nu} = 0.05$ to obtain acceptance rates close to 0.23, as recommanded by Gelman et *al*; m was set to 10 .

Annexe

C

Annexe concernant le chapitre 5

Supplementary materials: A Bayesian method to estimate the optimal threshold of a longitudinal biomarker

Fabien Subtil[1,2] and Muriel Rabilloud[1,2]

May 2009

[1] Université de Lyon, F-69000, Lyon; Université Lyon 1;
CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France.

[2] Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France.

1 Gibbs sampler for the simulation part

The model for the simulation part was:

$$\begin{aligned} &\log\left(y_{lj}\right) = \log\left(\exp\left(\eta_{1l}\right) + \exp\left(\eta_{2l}t_{lj}\right)\right) + \varepsilon_{lj} \\ &\varepsilon_{lj} \sim N\left(0, \sigma_{\varepsilon}^2\right) \quad \eta_{1l} \sim N\left(\mu_{\eta_1}, \sigma_{\eta_1}^2\right) \quad \eta_{2l} \sim 0.5 \times N\left(\mu_{\eta_2}, \sigma_{\eta_{20}}^2\right) + 0.5 \times N\left(\mu_{\eta_2} + \delta, \sigma_{\eta_{21}}^2\right) \end{aligned} \tag{1}$$

Let z_l denotes the group to which patient l belongs ($z_l = 0$ for patients belonging to the first group of the mixture for η_2 and z_l=1 for patients belonging to the second group of the mixture for η_2). Let n_0 denotes the number of patients belonging to the first group and n_1 the number of patients belonging to the second group. Let $\mathbf{y}$ denotes the data, m_l the number of measurements of patient l and m the total number of measurements.

The following notations were used for the hyperpriors:

$$\mu_{\eta_1} \hookrightarrow \mathcal{N}(\mu_{\mu_{\eta_1}}, \sigma^2_{\mu_{\eta_1}}) \quad \mu_{\eta_2} \hookrightarrow \mathcal{N}(\mu_{\mu_{\eta_2}}, \sigma^2_{\mu_{\eta_2}}) \quad \mu_{\delta} \hookrightarrow \mathcal{N}(\mu_{\mu_{\delta}}, \sigma^2_{\mu_{\delta}}) \quad z_l \hookrightarrow \mathcal{B}(1, 1/2)$$

$$1/\sigma_{\varepsilon}^2 \hookrightarrow \text{Gamma}(a_{\varepsilon}, b_{\varepsilon}) \quad 1/\sigma_{\eta_1}^2 \hookrightarrow \text{Gamma}(a_{\eta_1}, b_{\eta_1}) \quad 1/\sigma_{\eta_{20}}^2 \hookrightarrow \text{Gamma}(a_{\eta_{20}}, b_{\eta_{20}}) \quad 1/\sigma_{\eta_{21}}^2 \hookrightarrow \text{Gamma}(a_{\eta_{21}}, b_{\eta_{21}})$$

Let:

$$\psi(t_{lj}) = \ln(\exp(\eta_{1l} + \exp(\eta_{2l}t_{lj})))$$

The Gibbs sampler was used to alternately sample from the conditional posterior distribution of each parameter of the model. At each iteration of the Gibbs sampler, we proceed as follows:

1. Sample $[1/\sigma^2_\varepsilon | rest, \mathbf{y}]$ (where $rest$ denotes the other parameters of the model) from:

$$\text{Gamma}\left(\frac{m + a_\varepsilon}{2}, \frac{\sum_{l=1}^{L}\sum_{j=1}^{m_j}(\ln(y_{lj}) - \psi(t_{lj}))^2 + b_\varepsilon}{2}\right)$$

2. Sample $[\eta_{1l} | rest, \mathbf{y}]$ using a Metropolis with normal proposal, centered around the value of η_{1l} at the previous iteration. The log posterior $[\eta_{1l} | rest, \mathbf{y}]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma^2_\varepsilon}\sum_{j=1}^{m_j}\left(\psi^2(t_{lj}) - 2\ln(y_{lj})\psi(t_{lj})\right) + \frac{\eta^2_{1l} - 2\eta_{1l}\mu_{\eta_1}}{\sigma^2_{\eta_1}}\right]$$

3. Sample $[\eta_{2l} | rest, \mathbf{y}]$ using a Metropolis with normal proposal, centered around the value of η_{2l} at the previous iteration. The log posterior $[\eta_{2l} | rest, \mathbf{y}]$ is given by:

$$-\frac{1}{2}\left[\frac{1}{\sigma^2_\varepsilon}\sum_{j=1}^{m_j}\left(\psi^2(t_{lj}) - 2\ln(y_{lj})\psi(t_{lj})\right) + I[z_l = 0]\frac{\eta^2_{2l} - 2\eta_{2l}\mu_{\eta_2}}{\sigma^2_{\eta_{20}}} + I[z_l = 1]\frac{\eta^2_{2l} - 2\eta_{2l}(\mu_{\eta_2} + \delta)}{\sigma^2_{\eta_{21}}}\right]$$

where $I[z_l = 0] = 1$ if $z_l = 0$, and 0 otherwise.

4. Sample $[z_l | rest, \mathbf{y}]$ from $\mathcal{B}(1, p_1/(p_0 + p_1))$, where:

$$p_0 = \frac{1}{\sigma_{\eta_{20}}\sqrt{2\pi}}\exp\left(-\frac{(\eta_{2l} - \mu_{\eta_2})^2}{2\sigma^2_{\eta_{20}}}\right) \qquad \text{and} \qquad p_1 = \frac{1}{\sigma_{\eta_{21}}\sqrt{2\pi}}\exp\left(-\frac{(\eta_{2l} - \mu_{\eta_2} - \delta)^2}{2\sigma^2_{\eta_{21}}}\right)$$

5. Sample $[\mu_{\eta_1} | rest, \mathbf{y}]$ from:

$$\mathcal{N}\left(\frac{\sum_{l=1}^{L}\eta_{1l}/\sigma^2_{\eta_1} + \mu_{\mu_{\eta_1}}/\sigma_{\mu_{\eta_1}}}{N/\sigma^2_{\eta_1} + 1/\sigma^2_{\mu_{\eta_1}}}, \frac{1}{N/\sigma^2_{\eta_1} + 1/\sigma^2_{\mu_{\eta_1}}}\right)$$

6. Sample $[\mu_{\eta_2} | rest, \mathbf{y}]$ from:

$$\mathcal{N}\left(\left(I[z_l = 0]\frac{\sum_{l=1}^{L}\eta_{2l}}{\sigma^2_{\eta_{20}}} + \frac{\sum_{l=1}^{L}\eta_{2l} \times I[z_l = 1] - n_1\delta}{\sigma^2_{\eta_{21}}} + \frac{\mu_{\mu_{\eta_2}}}{\sigma^2_{\mu_{\eta_{20}}}}\right)/\tau_1, \frac{1}{\tau_1}\right)$$

where $\tau_1 = n_0/\sigma^2_{\eta_{20}} + n_1/\sigma^2_{\eta_{21}} + 1/\sigma^2_{\mu_{20}}$.

7. Sample $[\delta | rest, \mathbf{y}]$ from:

$$\mathcal{N}\left(\left(\frac{\sum_{l=1}^{L}\eta_{1l} \times I[z_l = 1] - n_1\mu_{\eta_2}}{\sigma^2_{\eta_{21}}} + \frac{\mu_\delta}{\sigma^2_\delta}\right)/\tau_2, \frac{1}{\tau_2}\right)$$

where $\tau_2 = n_1/\sigma^2_{\eta_{21}} + 1/\sigma^2_\delta$.

8. Sample $[1/\sigma^2_{\eta_1} | rest, \mathbf{y}]$ from:

$$\text{Gamma}\left(\frac{N + a_{\eta_1}}{2}, \frac{\sum_{l=1}^{L}(\eta_{1l} - \mu_{\eta_1})^2 + b_{\eta_1}}{2}\right)$$

9. Sample $[1/\sigma^2_{\eta_{20}}|rest, \mathbf{y}]$ from:

$$\text{Gamma}\left(\frac{n_0 + a_{\eta_{20}}}{2}, \frac{\sum_{l=1}^{L}(\eta_{2l} \times I[z_l = 0] - \mu_{\eta_2})^2 + b_{\eta_{20}}}{2}\right)$$

10. Sample $[1/\sigma^2_{\eta_{21}}|rest, \mathbf{y}]$ from:

$$\text{Gamma}\left(\frac{n_1 + a_{\eta_{21}}}{2}, \frac{\sum_{l=1}^{L}(\eta_{2l} \times I[z_l = 1] - \mu_{\eta_2} - \delta)^2 + b_{\eta_{21}}}{2}\right)$$

Glossaire

- ▷ **HDP** : région de plus haute densité. Pour un intervalle de crédibilité à 95 %, cette région contient 95 % de la densité a posteriori, la densité de probabilité des points lui appartenant étant au moins aussi élevée que celle des points ne lui appartenant pas.
- ▷ **UFHI** : ultrasons focalisés de haute intensité.
- ▷ **MCMC** : Monte Carlo Markov chain / chaîne de Markov - Monte Carlo.
- ▷ **Nadir** : plus basse valeur d'un ensemble de mesures.
- ▷ **Prévalence** : nombre de personnes atteintes d'une certaine maladie, à un moment donné et dans une population précise.
- ▷ **PSA** : prostate-specific antigen / antigène spécifique de la prostate.
- ▷ **ROC** : receiver operating characteristic curve. Courbe représentant la sensibilité d'un marqueur en fonction du complément de la spécificité, pour l'ensemble des seuils de positivité possibles. L'aire sous cette courbe est notée AROC.
- ▷ **Sen** : sensibilité. Probabilité qu'un test diagnostic soit positif pour une personne malade.
- ▷ **Spe** : spécificité. Probabilité qu'un test diagnostic soit négatif pour une personne non malade.
- ▷ **Vpp** : valeur prédictive positive. Probabilité qu'une personne soit malade sachant que le test diagnostic est positif.
- ▷ **Vpn** : valeur prédictive négative. Probabilité qu'une personne ne soit pas malade sachant que le test diagnostic est négatif.

Résumé

Lorsqu'un biomarqueur est mesuré de façon répétée au cours du suivi de patients, il est d'abord nécessaire d'établir un critère, issu du profil d'évolution longitudinal du marqueur, afin de détecter la survenue d'un événement, ou d'en prédire la gravité. Nous avons développé une méthode de modélisation robuste de données longitudinales, afin de calculer les différents critères pour les patients, et d'en comparer les performances diagnostiques ou pronostiques. Dans un second temps, il faut déterminer un seuil de ce critère quantitatif au dessus ou en dessous duquel le test diagnostique est considéré comme positif. Une méthode Bayésienne d'estimation de ce seuil et de son intervalle de crédibilité a été développée. Ce travail a été appliqué au diagnostic de persistance locale de cellules cancéreuses après traitement par ultrasons d'un cancer de la prostate. Ce diagnostic est effectué à partir des mesures répétées d'antigène spécifique de la prostate (PSA), dont le nadir a été retenu, avec différents seuils, comme meilleur critère diagnostique. Ceci permet de n'effectuer des biopsies que lorsqu'il y a de fortes chances qu'elles soient positives.

Mots-clés

Biomarqueur ; diagnostic précoce ; données longitudinales ; méthodes Bayésiennes semi-paramétriques ; seuil optimal ; antigène spécifique de la prostate.

Title

Methodology for the use of longitudinal quantitative biomarkers in medical decision making

Abstract

For the early diagnosis or prognosis of an event in presence of repeated measurements of a biomarker over time, it is necessary to define a criterion, stemming from the longitudinal profiles of that marker. A method was developed for a robust modelling of marker measurements, to calculate the various criteria for the patients, and compare their diagnostic or prognostic accuracies. Using the continuous criterion as a diagnostic test requires the specification of a threshold. A Bayesian method was developed to estimate this threshold and its credible interval. This method was applied to the diagnosis of local prostate cancer persistence after an ultrasound treatment. The diagnosis relies on serial measurements of prostate specific antigen (PSA), whose nadir (along with several thresholds) was found to be the best diagnostic criterion. This allows to trigger biopsy only when this biopsy is likely to be positive.

Keywords

Biomarker ; early diagnosis ; longitudinal study ; semi-parametric Bayesian methods ; optimal cut-point ; prostate specific antigen.

Intitulé et adresse du laboratoire

Laboratoire de Biométrie et Biologie Evolutive - Equipe Biostatistique-Santé
162, av Lacassagne
69003 Lyon France