Evaluation du risque de maladie : conception d'un processus et d'un système d'information permettant la construction d'un score de risque adapté au contexte, application au cancer du sein

Emilien Gauthier

N° d'ordre : 2013telb0268

Sous le sceau de l'Université européenne de Bretagne

Télécom Bretagne

En habilitation conjointe avec l'Université de Bretagne Sud

École Doctorale – SICMA

Évaluation du risque de maladie : conception d'un processus et d'un système d'information permettant la construction d'un score de risque adapté au contexte, application au cancer du sein

---

Thèse de Doctorat

Mention : « Sciences et Technologies de l'Information et de la Communication »

Présentée par **Émilien Gauthier**

Département : LUSSI
Laboratoire : Lab-STICC – Pôle : CID

Directeur de thèse : Philippe Lenca – Encadrants : Laurent Brisson, Jacques Simonin

Soutenue le 29 janvier 2013

Jury :

Rapporteurs :	Jacky Akoka	Professeur, Conservatoire National des Arts et Métiers
	Martine Collard	Professeur, Université des Antilles et de la Guyane
Examinateurs :	Giuseppe Berio	Professeur, Université de Bretagne Sud
	Laurent Brisson	Maître de conférences, Télécom Bretagne
	Thanh-Nghi Do	Maître de conférences, Université de Can Tho
	Philippe Lenca	Professeur, Télécom Bretagne
Invités :	Jacques Simonin	Directeur d'études, Télécom Bretagne
	Stéphane Ragusa	Maître de conférences détaché auprès de Statlife

ÉVALUATION DU RISQUE DE MALADIE : CONCEPTION D'UN PROCESSUS ET D'UN SYSTÈME D'INFORMATION PERMETTANT LA CONSTRUCTION D'UN SCORE DE RISQUE ADAPTÉ AU CONTEXTE, APPLICATION AU CANCER DU SEIN

Émilien Gauthier

Thèse

TELECOM Bretagne

Résumé

Résumé – Bien que de nombreux scores existent dans le domaine de la santé pour prédire un risque de maladie, ceux-ci sont peu utilisés alors qu'ils pourraient servir à individualiser la prévention pour la renforcer en fonction du niveau de risque estimé. Pour faciliter la production de scores performants dans la détection des profils à risque et adaptés au contexte d'utilisation, nous proposons un processus de construction de scores de risque.

Afin de mener des expérimentations, nous spécifions l'architecture d'un système d'information qui supporte les processus de production et d'utilisation de scores de risque. Grâce à la mise en œuvre d'une partie de cette architecture, nous utilisons notre processus pour expérimenter la création de scores de risque du cancer du sein basés sur une base de données américaine publique et sur les données françaises de l'étude de cohorte E3N.

Sur l'exemple du cancer du sein, nous montrons qu'il est possible d'obtenir des performances comparables en termes de discrimination et supérieures en termes de calibration à celles de la littérature avec l'algorithme des plus proches voisins qui est compréhensible par les médecins et patients, tout en utilisant moins d'attributs.
Mots clefs – Score de risque, cancer du sein, proches voisins, système d'information.

Title – Assessing disease risk : a process and information system design that support the construction of a risk score adapted to the context, application on breast cancer

Abstract – Although there are many risk scores in the health field to predict disease risk, they are not as used as they could be to individualize and enhance prevention based on an estimated risk level. In order to facilitate the production of risk scores that are efficient in detecting high risk profiles and that fit to the context of use, we suggest a risk score building process.

In order to conduct experiments, we build an information system architecture that supports the building and use process of risk scores. Thanks to the implementation of this architecture, we use our process to experiment the creation of breast cancer risk scores based on a publicly available american database and on the E3N French cohort study database.

Using the breast cancer example, we show that it is possible to obtain comparable performances in terms of discrimination and better performances in calibration than available risk scores of the literature, using a readable k-nearest-neighbor algorithm and less attributes.
Keywords – Risk score, breast cancer, k-nearest-neighbor, information system.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Philippe Lenca, et mes encadrants, Laurent Brisson et Jacques Simonin, qui ont rendu possible le déroulement de cette thèse de fouille de données à la frontière de l'épidémiologie. Un merci particulier à Laurent dont le suivi et les conseils ont été appréciés malgré la distance entre Brest et Villejuif.

Je souhaite remercier Stéphane Ragusa qui a permis la poursuite de mon parcours vers la thèse en proposant à la fois une thématique intéressante et en permettant le financement de mon travail par Statlife avec l'aide de l'Association Nationale de la Recherche et de la Technologie.

Je tiens à témoigner ma gratitude aux Professeurs Martine Collard et Jacky Akoka pour avoir accepté la charge de rapporteur de cette thèse. Je suis également reconnaissant à Giuseppe Berio et Thanh-Nghi Do d'avoir participé à l'évaluation de mon travail.

Merci à Françoise Clavel-Chapelon qui m'a accueilli depuis 2007 comme un membre à part entière de son équipe E3N « Nutrition, Hormones et Santé des Femmes » de l'Institut National de la Santé et de la Recherche Médicale au sein de l'Institut Gustave Roussy. Je remercie également Yvon Kermarrec qui a permis mon intégration au département LUSSI de Télécom Bretagne lors de mes passages à Brest.

Merci également à ceux qui se sont intéressés de près ou de loin à mon travail, Gilles Coppin, Patrick Meyer, Guy Fagherazzi, Philippe Picouet en me faisant part de leurs idées ou en me prodiguant des conseils.

Merci à Maryvonne Niravong et Ghislaine Le Gall, qui ont facilité les rapports avec les administrations. Plus généralement, merci à tous les membres qui font fonctionner au quotidien, et ce n'est pas une mince affaire, l'équipe E3N, le département LUSSI et la société Statlife, des collègues avec qui le travail a toujours été très agréable.

Merci à Lyan Hoang de m'avoir facilité l'accès aux données de l'équipe à Villejuif. Merci à Sébastien Bigaret et Patrick Meyer qui ont permis la mise à disposition du serveur de calcul Éole à Télécom Bretagne, et en particulier à Sébastien dont la promptitude du règlement des problèmes d'accès au serveur depuis Villejuif et les conseils en optimisation ont été utiles et très appréciés.

Merci à tous les collègues qui ont contribué à maintenir une bonne ambiance de travail dans les bureaux successifs où j'ai posé mon ordinateur et qui ont toujours répondu à mes questions en épidémiologie, en statistique ou en informatique : Céline, Marina, Pierre, Marie-Noël, Lionel, Laura, Alban, Anne, Alice, Laureen et Suvi à Villejuif ; Cyril, Sahbi, Wided, Pierre, Santiago, Thomas, Yannick pour son travail sur le thème LaTeX utilisé dans ce manuscrit, Clément, Émile, Vanea, Nyana, Iyas,

Benoit et Romain à Brest. Merci également à mes prédécesseurs de doctorants dont je n'ai pas eu la chance de partager le bureau, mais qui ont toujours été de bon conseil, dont Jérôme, Guy, Marina, Gaël, Agnès et Énora.

Je double les remerciements à Céline, Alice, Anne, Laureen, Guy, Agnès, Julien, Laura, mes parents et mes encadrants au sens large qui ont eu le courage et la gentillesse de relire dans le détail tout ou partie de ce manuscrit.

Une pensée pour ma famille en général, en particulier pour mes parents, pour Clélia et Ludiane, et pour les amis qui ont été des soutiens indispensables au cours de ces trois années de thèse, malgré les hauts et les bas. Un tendre merci à Céline qui m'a supporté au quotidien dans cette réalisation.

Table des matières

Liste des tableaux

Table des figures

Introduction

La prévention des grandes maladies

La lutte contre les grandes maladies telles que les cancers, les maladies cardio-vasculaires ou le déclin cognitif se répartit globalement suivant trois grands axes. Le premier axe consiste en l'amélioration des traitements utilisés une fois la maladie détectée. Le deuxième axe consiste en la détection de la maladie au plus tôt, de nombreuses études montrant qu'une prise en charge précoce de la maladie réduit les taux de mortalité. Enfin, le troisième axe consiste en la mise en œuvre de mesures permettant d'éviter d'être atteint par la maladie. Si ces axes bénéficient de toute l'attention des acteurs du monde de la santé, nous pensons que le dépistage et la prévention des grandes maladies peuvent être encore améliorés afin de diminuer le nombre de nouveaux cas ou la gravité de ces grandes maladies.

L'amélioration du dépistage et de la prévention passe par l'augmentation des connaissances sur les maladies, notamment leurs aspects épidémiologiques* [1] et étiologiques*. C'est le travail que mènent les chercheurs en santé publique à l'aide, par exemple, de grandes enquêtes de cohorte* qui permettent de suivre l'évolution de la santé de personnes sur de longues périodes pour quantifier l'impact de différents facteurs de risque sur les maladies.

Mais l'amélioration du dépistage et de la prévention passe également par une meilleure communication des connaissances scientifiques vers les médecins et les patients. D'une part, nous émettons l'hypothèse que l'utilisation d'un score de risque peut permettre d'améliorer le dépistage en proposant un suivi adapté au niveau de risque estimé pour la personne. Par exemple, des mammographies* de contrôle pourraient avoir lieu plus tôt ou plus tard dans la vie d'une femme en fonction de son niveau de risque. D'autre part, ces scores pourraient améliorer la prévention en population générale en permettant à chacun, non seulement d'estimer son risque pour une maladie, mais surtout de comprendre facilement l'impact des différents facteurs de risque sur le risque estimé et, par conséquent, de mieux appréhender l'influence de différentes modifications comportementales sur ce risque. Par exemple, les utilisateurs de tels scores de risque pourraient avoir des réponses aux questions comme : l'activité physique diminue-t-elle le risque de maladie cardio-vasculaire ? Et, si oui, dans quelles proportions ?

Nos objectifs

Afin d'améliorer la prévention en santé publique grâce à l'utilisation de scores de risque, nous avons défini trois objectifs majeurs que sont la mise au point d'un processus de création de scores de risque qui permette d'adapter le score à son contexte d'utilisation, la conception d'un système d'information destiné à soutenir

1. Un astérisque symbolise un terme présent dans le glossaire page 153.

la réalisation d'expérimentations qui s'appuient sur le processus et l'utilisation d'un modèle d'évaluation à la fois compréhensible et performant dans la détection des profils d'individus à risque en population générale.

L'objectif de simplification et d'adaptation du processus de création de scores de risque doit permettre de faciliter la création de score de risque qui soient adaptés au contexte d'utilisation. En effet, pour le cancer du sein par exemple, un score de risque ne devra pas tirer parti des mêmes attributs selon qu'il soit conçu pour être utilisé par un patient, un médecin généraliste ou un radiologue. Le radiologue disposant d'éléments comme la densité mammaire dont ne dispose pas forcément le médecin et le patient n'ayant pas les mêmes moyens d'agir sur le risque qu'un médecin. Afin de pouvoir être utilisé sur des maladies différentes, le processus doit être générique et pas seulement adapté au cancer du sein, maladie sur laquelle nous appliquons nos propositions.

L'objectif de soutenir le processus proposé par un système d'information implique l'automatisation du déroulement des expérimentations qui conduisent à la création et à l'utilisation d'un score de risque. Le système d'information doit être générique du point de vue des maladies. Il doit être caractérisé par une vue fonctionnelle qui soit utilisable pour les développements de systèmes informatiques de ce système d'information. L'utilisation de ce système d'information doit déboucher sur la proposition d'un score de risque adapté au contexte de la prévention du cancer du sein dans une clinique du risque et sur la mise en œuvre de ce score dans un prototype de système informatique permettant l'utilisation du score de risque.

Enfin, nous avons défini l'objectif d'utiliser un modèle d'évaluation du risque qui soit à la fois plus compréhensible que les modèles utilisés pour construire les scores existants, mais au moins aussi performant en matière de détection des profils à risque de maladie.

Nos contributions

Pour répondre à l'objectif de création d'un score de risque adapté au contexte d'utilisation, nous proposons de mener un travail à dominante informatique qui soit à la frontière de l'épidémiologie pour tenter d'améliorer les scores produits à base de méthodes statistiques.

En cela, nous commençons par proposer un processus de construction de score de risque de maladie, basé sur l'exemple du cancer du sein, qui permet la réalisation d'un compromis entre la performance du score à évaluer correctement un niveau de risque, l'adaptation des attributs utilisés par rapport au contexte d'utilisation et la prise en compte des coûts de différentes natures pour les attributs retenus.

Afin de répondre à l'objectif d'automatisation du processus de création de scores de risque, nous proposons l'architecture d'un système d'information qui supporte les processus métiers de production et d'utilisation de scores de risque. Nous mettons en œuvre des fonctions extraites de cette architecture sous la forme d'un système informatique (de l'architecture applicative au code) qui tient compte des contraintes métier.

Enfin, pour répondre à l'objectif d'utiliser un modèle d'évaluation compréhensible tout en maintenant des performances au moins équivalentes à celles de la littérature, nous proposons l'utilisation de l'algorithme des plus proches voisins pour produire un score aisément explicable à l'utilisateur et dont nous avons mesuré, pour le cancer du sein, les performances sur un jeu de données américaines publiques et un jeu de données françaises afin de permettre l'utilisation du score en France.

L'organisation du mémoire

Les problématiques soulevées, les objectifs fixés et les solutions que nous proposons s'articulent de la manière suivante dans le manuscrit.

Dans le **chapitre 1**, nous détaillons les différents types de prévention pour les maladies en général et la manière dont les programmes sont menés en France. Nous avons choisi d'utiliser nos propositions sur le cas du cancer du sein, en conséquence nous présentons des éléments d'épidémiologie* et de physiologie de cette maladie avant d'aborder le détail des facteurs de risque connus.

Grâce à des exemples représentatifs, nous réalisons dans le **chapitre 2** un état de l'art des méthodes disponibles dans la littérature pour construire un score de risque de maladie. Nous détaillons également les méthodes de mesure de la performance des scores de risque conçus, que ce soit en termes de discrimination ou de calibration.

En prenant en compte les faiblesses que nous constatons sur les scores actuels, nous proposons dans le **chapitre 3** un processus de construction d'un score de risque qui permet d'intégrer les connaissances des épidémiologistes du domaine et le contexte d'utilisation du score au travers des besoins des utilisateurs. Un exemple de conception d'un score de risque pour le cancer du sein est présenté en fin de chapitre.

Afin de faciliter les expérimentations basées sur le processus proposé, nous décrivons dans le **chapitre 4** une proposition d'architecture de système d'information du risque de maladie qui supporte les processus de production et d'utilisation de scores de risque. Nous détaillons l'architecture fonctionnelle et nous expliquons la mise en œuvre des fonctions extraites de cette architecture sous la forme d'un système informatique qui tient compte des contraintes métier.

Nous montrons dans le **chapitre 5** les résultats obtenus en termes de modélisation du risque de cancer du sein. Pour permettre la comparaison de la performance de notre algorithme de type proches voisins, nous construisons un score de risque à l'aide d'une base publique américaine. Puis, afin de produire un score de risque adapté à la population française, nous utilisons les données de la cohorte française E3N pour construire un score de risque dont la performance est évaluée au moyen de mesures de la calibration et de la discrimination.

1 La prévention des grandes maladies

Si la lutte contre les maladies se fait essentiellement en cherchant et en améliorant les traitements utilisés une fois la maladie déclarée, il existe toutefois un autre moyen de lutte qui est d'améliorer la prévention, une suite d'actions qui peuvent se dérouler avant que la maladie soit déclarée pour empêcher ou ralentir la survenue de celle-ci.

Dans ce chapitre, nous expliquons les différents types de prévention qui existent pour les grandes maladies et nous analysons la manière dont elles peuvent bénéficier des scores de risque dans le cadre d'une politique de santé misant sur la prévention, la prédiction, la personnalisation et la participation. Nous détaillons ensuite, du point de vue épidémiologique et physiologique, le cas du cancer du sein sur lequel nous appliquons nos travaux.

1.1 Prévention : les concepts

Dans cette première partie, nous explicitons les différents niveaux de prévention, l'intérêt de la prévention personnalisée et ses enjeux au travers de l'exemple d'une clinique du risque pour le cancer du sein.

1.1.1 Définitions et histoire de la prévention

On définit la prévention comme un « ensemble de mesures destinées à éviter un événement qu'on peut prévoir et dont on pense qu'il entraînerait un dommage pour l'individu ou la collectivité » [Imbs 94]. Appliqué au domaine de la santé publique, le concept de prévention est appelé la *prophylaxie*, elle s'oppose au concept de la guérison, car son objectif est de devancer la maladie plutôt que d'agir pour essayer de l'éliminer.

Historiquement, l'Organisation Mondiale de la Santé (OMS), à partir de travaux d'une commission américaine [USC 57], propose de définir trois niveaux principaux de prévention en fonction du fait que la maladie soit absente ou présente chez le patient d'une part et en fonction du niveau de risque que la maladie soit présente d'autre part.

- La prévention primaire, lorsque la maladie est considérée comme absente et que le risque d'avoir la maladie est faible, regroupe les mesures qui sont utilisées pour tenter d'éviter la survenue de la maladie. Les programmes de prévention

primaire sont généralement diffusés à l'échelle d'une population et incluent, par exemple, la vaccination ou des conseils sur le plan de la nutrition ou de l'activité physique (Plan National Nutrition Santé en France dont le slogan était « 5 fruits et légumes par jour »).

- La prévention secondaire, lorsque la maladie est considérée comme absente, mais que le risque d'avoir la maladie est élevé, regroupe les mesures qui sont utilisées pour diagnostiquer la maladie au plus tôt afin d'augmenter la probabilité de guérison. Ce type de prévention inclut essentiellement les programmes de dépistage au sein de larges sous-populations à risque, par exemple le dépistage du cancer du sein après 50 ans en France ou du cancer colorectal.
- La prévention tertiaire, lorsque la maladie est considérée comme présente et que la maladie effectivement est présente, regroupe les mesures qui sont utilisées pour diminuer les complications dues à la maladie et la rechute.
- Enfin, la prévention quaternaire, qui ne fait pas partie de la classification de l'OMS, lorsque le patient est considéré comme malade, mais que la maladie est absente, est apparue plus récemment et regroupe les mesures qui visent à limiter la surmédicalisation.

Mais plus récemment, les notions de prévention primaire et secondaire ont eu tendance à être remplacées par les notions de prévention universelle (ou généralisée), prévention sélective et prévention indiquée [INSERM 09] [1]. Les interventions de prévention universelle sont conçues pour des sous-groupes de la population générale qui n'ont pas été sélectionnés par rapport à un risque défini (campagne de vaccination en milieu scolaire par exemple). La prévention sélective vise, elle, plus spécifiquement un sous-groupe de sujets qui présente un risque significativement plus élevé de développer une maladie. Enfin, la prévention indiquée regroupe les actions menées pour des sujets qui ont des *signes d'appel* pour une maladie sans toutefois utiliser des critères diagnostiques.

Cette nouvelle manière de découper les différentes actions possibles dans le domaine de la prévention met clairement en avant la notion de sélection des sous-populations à risque auxquelles devront s'adresser en priorité les actions de prévention. Il faut donc être capable de sélectionner au mieux des sous-populations. C'est dans le contexte de prévention sélective que se situent ces travaux de thèse. La sélection des sous-populations est en partie le but de l'épidémiologie* dont l'objet est la « quantification de la fréquence d'un événement de santé dans une population et la détermination de ses causes biologiques et médicales, environnementales, socio-économiques, etc ». Son objectif final est d'identifier les facteurs qui causent la survenue d'événements de santé pour pouvoir les éliminer ou les limiter. C'est de cette discipline que les premiers scores de risque, utilisés à grande échelle dans le domaine médical, sont nés (voir l'origine des scores de risque en santé, partie 2.1, page 23).

1. INSERM : Institut National de la Santé et de la Recherche Médicale

Au niveau mondial, c'est l'OMS et sa conférence sur la promotion de la santé à Ottawa en 1986 qui ont marqué le point de départ de la prise en compte des problématiques globales de prévention dans les politiques nationales de santé publique. Ces politiques étaient jusque-là largement focalisées sur la médecine curative. En France, ce sont diverses agences, nées de la volonté de structuration de la politique de santé, qui ont hérité de la définition de différentes composantes de la prévention dans tous les domaines qui étaient susceptibles de provoquer de la surmortalité. Le Haut Comité de Santé Publique créé 1991 et son successeur le Haut Conseil de la Santé créé en 2002 ont, par exemple, pour mission d'orienter la politique de santé française au travers de rapports thématiques triennaux, notamment en ce qui concerne la prévention. Mais si l'orientation se décide au niveau national, les collectivités locales ont en charge la mise au point de programmes de dépistage qui sont une composante de la prévention (secondaire ou sélective). Par exemple, les départements ont la charge, depuis une loi de décentralisation de 1983, des dispensaires contre la tuberculose ou encore des campagnes de vaccination. Ils avaient déjà la responsabilité de la mise en place des programmes de dépistage des affections cancéreuses depuis 1963, sous la responsabilité de l'État et le financement partiel de l'Assurance Maladie [IGAS 03].

Depuis, les différents plans nationaux et les rapports thématiques, chargés de définir des programmes de lutte contre certaines maladies parmi les plus répandues ou de faire le bilan d'une action, ont intégré pour nombre d'entre eux un axe dédié à la prévention. Par exemple, le plan Alzheimer 2008-2012 prévoyait trois mesures pour informer et sensibiliser le grand public. De même, le plan cancer 2009-2013 publié par l'Institut National du Cancer (INCa) inclut huit mesures dans un axe consacré à la prévention et au dépistage [INCa 09]. Ces mesures prévoient notamment la lutte contre les inégalités d'accès et de recours au dépistage en plaçant le « médecin traitant au cœur du dispositif de dépistage » ou encore l'amélioration du dispositif des programmes nationaux de dépistage organisé des cancers.

Une agence s'est intéressée, de manière plus spécifique, à la prévention du risque cardio-vasculaire : dans un rapport publié en 2004 [ANAÉS 04], l'Agence Nationale d'Accréditation et d'Évaluation en Santé (ANAÉS) analyse les méthodes d'évaluation du risque cardio-vasculaire global. L'objectif de prévention primaire est alors manifeste : prédire le risque cardio-vasculaire global pour « les sujets n'ayant aucune pathologie cardio-vasculaire cliniquement exprimée et pour lesquels le dépistage et la prise en charge permettraient d'éviter, limiter ou retarder le développement d'une maladie cardio-vasculaire ». Dans ce rapport sont abordées des questions précises comme la pertinence de la démarche d'utilisation de score, les types de modèles à utiliser pour prédire les risques et leur adaptation à la population française ou encore les bénéfices attendus de l'utilisation de scores de risque dans la prise en charge personnalisée des patients.

1.1.2 La prévention personnalisée

Dans cette partie, nous abordons la personnalisation de la prévention et de la médecine, le concept de clinique du risque et l'intérêt des acteurs de la recherche au travers de l'exemple de la Fondation ARC.

1.1.2.1 Personnalisation de la médecine

Si, au cours des vingt dernières années, de nombreux modèles ont été construits par les épidémiologistes et les statisticiens pour mesurer le risque de maladie (voir chapitre 2), il faut distinguer les utilisations qui peuvent en être faites et les enjeux associés. Premier type d'utilisation, la modélisation du risque en fonction de facteurs de risque permet aux décideurs (gouvernement, agences de santé, Assurance Maladie, mutuelles, etc.) d'effectuer des analyses de coûts-bénéfices au niveau d'une population. L'enjeu est la prise de décisions globales quant au risque d'une maladie. Par exemple : comment fixer l'âge à partir duquel les femmes sont encouragées à faire régulièrement des mammographies de contrôle ? À quel intervalle de temps correspond le *régulièrement* ? En revanche, second type d'utilisation, à destination des médecins cette fois, la modélisation du risque doit permettre une évaluation fiable des risques d'un patient en particulier en ce qui concerne la sensibilité (la capacité à donner un résultat positif lorsque la maladie est présente) et la spécificité (la capacité à donner un résultat négatif quand la maladie est absente) afin d'appuyer une décision. L'enjeu est alors d'aider le médecin à évaluer une situation pour décider du type de prise en charge (simple conseil, examen, médication, etc.) en fonction du niveau estimé de risque, c'est la prévention personnalisée.

La prévention personnalisée, c'est aussi le pendant d'une médecine personnalisée fondée sur des preuves. Selon [Sackett 96], la médecine fondée sur des preuves se base sur l'expérience clinique, le patient et les données de la recherche. Elle est le résultat d'un processus qui débute par la formulation d'une question clinique. Elle se poursuit par une phase de recherche dans la littérature médicale et l'évaluation de la validité des éléments extraits. En fonction de leur validité, ces éléments peuvent être utilisés dans la pratique médicale en fonction de la situation du patient, de son état clinique et de ses préférences.

La médecine personnalisée prolonge la logique de la médecine fondée sur les preuves en décrivant la prise en charge d'un patient en fonction de ses caractéristiques génétiques et environnementales qui pourront modifier sa réponse à un traitement. Par exemple, un cancer du sein n'est plus soigné aujourd'hui par un traitement standard et appliqué à tous les cancers du sein. Différents types de cancers du sein existent que ce soit par leurs différents stades d'évolution, leurs différentes signatures génétiques ou encore leurs différentes réponses hormonales. Chaque traitement est adapté au type de cancer diagnostiqué et aux caractéristiques du patient. Ce concept, appliqué avant l'apparition ou le diagnostic d'une maladie, implique que la prévention peut être différente, évidemment en fonction de la maladie, de ses

caractéristiques, mais aussi des caractéristiques du patient.

1.1.2.2 Concept de clinique du risque

La médecine personnalisée est l'inspiration du projet de clinique du risque en cours à Villejuif (94) et couvrant plusieurs départements du sud de la région Île-de-France (3,8 millions d'habitants en 2009). Le projet est porté par l'Institut Gustave Roussy (IGR) considéré comme le plus grand centre de lutte contre le cancer en Europe. À ce projet sont associés plusieurs services de l'institut, trois laboratoires de la recherche publique (rattachés à l'INSERM ou au CNRS) et des associations de lutte contre le cancer. Cette clinique du risque a pour principal objectif de montrer qu'il est possible d'individualiser la gestion des risques de cancer du sein en diminuant les coûts et en augmentant l'espérance de vie théorique. Pour cela, il faut être capable d'évaluer les risques individuels de cancer du sein, de proposer une prise en charge individualisée et évolutive dans le temps en cas de risque augmenté puis d'améliorer ces deux étapes en identifiant des biomarqueurs prédictifs et en développant de nouvelles modalités de dépistage en biologie et en imagerie. Parmi les valorisations scientifiques possibles, on note la modification des pratiques de prévention avec la proposition de nouveaux modèles de prise en charge des risques, l'exportation du modèle de clinique du risque à d'autres maladies et d'autres pays, le développement de molécules de chimioprévention ou encore l'élaboration de nouveaux outils de calcul du risque.

Cette clinique du risque fonctionnera en deux étapes décrites ci-dessous.

- Une première étape sera l'évaluation du risque individuel en précisant le risque de cancer du sein afin de proposer une prise en charge adaptée au niveau de risque. Par exemple, dans le cas d'un risque élevé, une proposition de programme de dépistage et de prévention intensifiée dans le cadre du réseau avec une consultation initiale à l'IGR pour information et prélèvement est faite. Dans le cas d'un risque intermédiaire, une proposition de surveillance auprès des gynécologues et radiologues du réseau régional de la clinique du risque est faite. Ou, dans le cas d'un risque moyen, une proposition de surveillance selon les modalités de la population générale est faite. Le risque individuel sera évalué par le médecin grâce à l'utilisation d'un questionnaire permettant d'estimer un niveau de risque. Cette phase permettra la constitution et la mise à jour d'une base de données de cohorte permettant d'améliorer l'évaluation du risque dans le futur.
- Une seconde étape est la prise en charge en fonction du risque. Si les femmes à risque intermédiaire et moyen seront suivies dans le réseau de la clinique du risque, en revanche les femmes à haut risque se voient déjà proposer à l'heure actuelle une consultation unique réalisée sur une seule journée à l'IGR. Cette journée permet un recueil complet des données cliniques et d'anamnèse*, une prise de sang pour mise en banque et utilisation future, une information complète de la patiente sur les risques et les modalités de prise en charge,

une imagerie initiale (mammographie, échographie ou imagerie par résonnance magnétique, IRM) plus, éventuellement d'autres examens de routine ou de recherche. Un programme personnalisé de surveillance et d'orientation vers une prise en charge dans le réseau de soin sont remis à la femme. Bien que déjà en partie existante, cette phase pourrait bénéficier de la mise en place d'outils d'information supplémentaires sur le risque de cancer du sein.

Plus spécifiquement, la méthode d'estimation du risque de cancer du sein utilisée dans cette clinique du risque devra être mise au point sur une cohorte (décrite en partie 3.2.1.2, page 66) de l'INSERM du fait des difficultés à transposer les modèles conçus dans d'autres pays vers la France et de la difficulté à ajouter d'autres facteurs de risque que ceux prévus initialement. La méthode d'estimation du risque choisie pour cette clinique du risque doit notamment être flexible pour permettre l'ajout de facteurs au cours de la vie du projet et intégrer des données d'exposition récentes. Le chapitre 5 de ce travail de thèse constitue une proposition de méthode d'estimation du risque pour la clinique du risque, sans préjuger des facteurs de risque qui seront spécifiquement retenus.

1.1.2.3 Médecine 4P : prévention, prédiction, personnalisation et participation

À un niveau supérieur, la clinique du risque s'insère dans un système de santé qui commence à se restructurer. À ce titre, l'évolution de l'ARC (Association pour la Recherche sur le Cancer) en Fondation ARC pour la recherche sur le cancer est intéressante. Annoncée en octobre 2012, cette restructuration se fonde sur deux axes principaux : l'ambition de guérir deux cancers sur trois en 2025 et d'investir le champ de la médecine 4P, pour prévention, prédiction, personnalisation et participation. Le financement de la clinique du risque est l'une des premières réalisations concrètes de cette politique.

L'objectif de cette médecine 4P est d'aller plus loin que la simple évolution vers une médecine personnalisée. Cette volonté est issue de l'observation que dans les pays développés, une part supérieure à 75 % des dépenses est consacrée aux maladies chroniques comme les maladies cardio-vasculaires, les cancers ou le diabète, des maladies dont il peut exister un grand nombre de sous-types qui peuvent chacun dépendre de dizaines de voies moléculaires différentes, chacune impliquant plusieurs centaines de molécules différentes. Ainsi, la médecine 4P se fonde sur une approche de précision axée sur les différences individuelles. La prévention vise à proposer des mesures correctives avant l'apparition de la maladie, la prédiction vise à évaluer le risque d'un individu pour une maladie, la personnalisation doit permettre d'adapter le traitement le plus finement possible à la maladie et à l'individu et la participation consiste à rendre l'individu acteur de sa santé, avant et après l'apparition de la maladie [Fondation ARC 12].

La mise en place d'une telle médecine implique, selon la Fondation ARC, l'augmentation de l'effort en matière de recherche et la mise en commun des savoirs qu'ils concernent la génétique, l'oncologie, la diététique, la virologie, la psychologie

ou les sciences sociales. La clinique du risque évoquée précédemment est un premier exemple de mise en commun de ces savoirs avec, en plus, une unité de lieu pour faciliter la mise en place du projet.

Globalement, les enjeux principaux de la prévention sont donc de diminuer l'incidence* des maladies puis, une fois la maladie déclarée, de diminuer le temps avant son diagnostic pour augmenter les chances de guérison. Une réponse possible est la mise en place de programmes de prévention personnalisée permettant de consacrer l'utilisation des moyens à disposition aux personnes les plus à risque tout en augmentant leur efficacité. Le début de promotion de la médecine 4P par la Fondation ARC et le lancement de la clinique du risque constituent un excellent cadre d'application pour les travaux présentés dans cette thèse. Le cancer du sein étant l'une des grandes maladies les plus fréquentes en France, nous avons choisi de l'utiliser comme support à l'application du processus et à la méthode d'évaluation du risque que nous proposons.

1.2 Le cas du cancer du sein

Parmi les grandes maladies, le cancer du sein tient une place importante dans le paysage médical français, notamment à cause du nombre élevé de nouveaux cas annuels. Nous avons choisi d'appliquer nos travaux à cette maladie, car elle concentre un grand nombre des défis qui peuvent être rencontrés dans le cadre de la création de scores de risque dans le domaine médical. Si, comme pour la majorité des grandes maladies, le nombre de nouveaux cas reste faible à l'échelle d'une population, en revanche les causes du cancer du sein ne sont pas toutes identifiées et les facteurs de risque connus peuvent difficilement être modifiés pour influer sur le risque. Cependant, il existe, pour le cancer du sein, des bases de données de qualité qui permettent d'étudier la maladie, ce qui n'est pas le cas pour toutes les grandes maladies. Il est donc nécessaire de mieux connaître ce cancer avant de l'utiliser comme cas d'application dans la suite du manuscrit. Dans cette sous-partie, nous abordons l'épidémiologie du cancer du sein, puis l'aspect physiologique de ce cancer à travers l'évocation des bases de la cancérogenèse et enfin, ses principaux facteurs de risque.

1.2.1 Épidémiologie du cancer du sein

Au niveau mondial, le nombre de cas de cancer et de décès dus au cancer augmente, notamment à cause de la hausse de la population et de sa moyenne d'âge et en particulier à cause des comportements à risque de ces populations.

État des lieux : Selon les dernières estimations du Centre International de Recherche sur le Cancer (CIRC) datant de 2008 [Ferlay 10], cette année-là, 12,7 millions de cas de cancer ont été recensés au niveau mondial pour environ 7,6 millions de décès dus au cancer. Parmi eux, le cancer du sein est le plus fréquemment diagnostiqué puisqu'il compte pour 23 % de l'ensemble des cancers diagnostiqués, soit 1,38 million de cas. Il est la cause de décès par cancer la plus courante chez les femmes avec 14 % des cas de décès par cancer, soit 0,46 million de cas [Jemal 11].

En France, la mortalité par cancer est codée et enregistrée au niveau national par le Centre d'épidémiologie sur les causes médicales de décès (CépiDC) grâce aux certificats de décès. Toutefois, tout comme au niveau mondial, l'incidence* du cancer en France reste difficilement mesurable. En revanche, elle est estimée à partir de registres généraux au niveau départemental et de registres spécifiques de certaines localisations de cancer créés à partir de 1975. Depuis 1991, ces registres sont regroupés dans le réseau FRANce - Cancer - Incidence et Mortalité (FRANCIM) qui couvre environ 13 % de la population française.

En France, en 2008, l'incidence estimée était de 99,7 nouveaux cas de cancer du sein par an pour 100 000 femmes (voir figure 1.1) selon [Ferlay 10], tandis que le taux de mortalité s'élevait à 17,6 décès par cancer du sein par an pour 100 000 personnes, ce qui représente une estimation de 51 012 nouveaux cas de cancer par

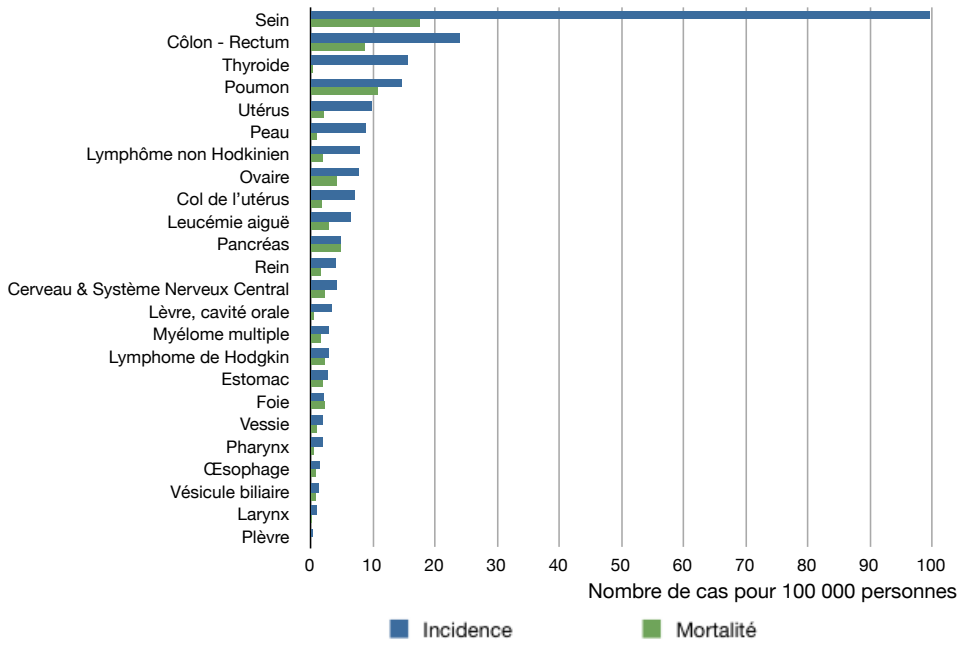

Figure 1.1 : Taux d'incidence et de mortalité annuelle, standardisé par âge, pour 100 000 personnes-années en France en 2008 selon [Ferlay 10].

an et 11 773 décès par an. Représentant 35,5 % des cancers féminins, le cancer du sein est le plus léthal chez la femme, suivi du cancer du côlon-rectum (12,8 %) et du poumon (5,7 %).

Évolution de l'incidence : Depuis 1980, année depuis laquelle l'incidence et la mortalité sont mesurées de manière standardisée, on constate que l'incidence est en forte hausse en France (voir figure 1.2) passant de 56,8 cas de cancer du sein pour 100 000 femmes par an en 1980 à 101,5 cas en 2005 selon [Belot 08]. À noter que ce chiffre est en légère baisse dans les données 2008 du CIRC publiées par [Ferlay 10].

La mortalité liée au cancer du sein en revanche, après une légère hausse entre 1980 et 1990, est en baisse sensible depuis lors, passant de 20,1 à 17,7 décès par an pour 100 000 femmes en 2005, soit une baisse de 12 % sur 15 ans. Plusieurs facteurs peuvent expliquer le contraste entre la nette hausse de l'incidence et la baisse sensible de la mortalité. Parmi eux, on relèvera la prise de conscience de l'existence de facteurs de risque entraînant des changements de style de vie, l'amélioration du dépistage par la mise en place de programmes à l'échelle nationale conduisant à un diagnostic plus précoce (et entraînant une hausse transitoire de l'incidence) ou encore l'amélioration des traitements. Par ailleurs, la baisse rapportée par [Ferlay 10] par rapport aux données publiées par [Belot 08] pourrait être expliquée par une baisse de 62 % de la consommation de traitements hormonaux de la ménopause entre 2000 et 2006 [Ringa 08] (voir l'impact de l'imprégnation hormonale page 19).

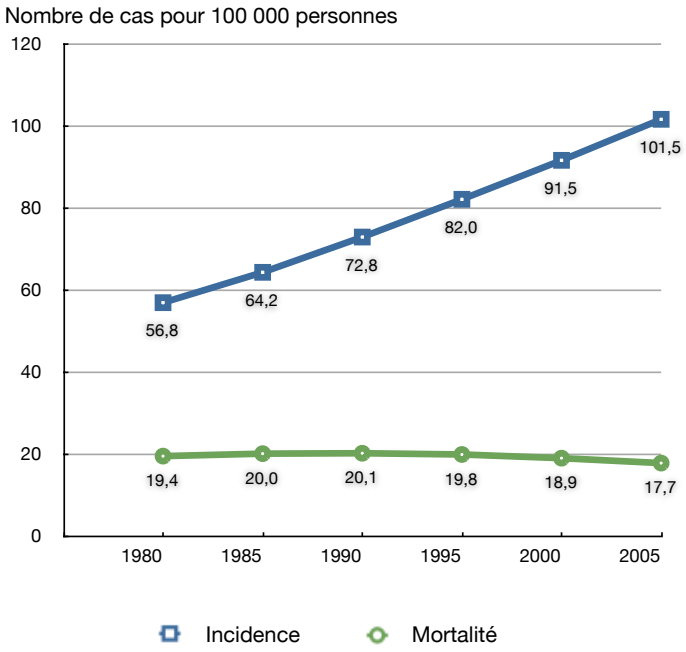

Figure 1.2 : Évolution des taux d'incidence et de mortalité annuelle, standardisés par âge, pour 100 000 personnes-années pour le cancer du sein de 1980 à 2005 en France selon [Ferlay 10].

Effet cohorte : Il est important de noter que l'incidence du cancer du sein est fortement associée à un effet cohorte. Le risque de développer un cancer du sein est en effet fortement lié à l'année de naissance de la femme : au même âge, deux femmes ayant des années de naissance différentes n'auront pas le même risque de cancer du sein. Par exemple, le risque de cancer du sein d'une femme née en 1910 sera deux fois moindre que celui d'une femme née en 1930 et presque trois fois moindre que celui d'une femme née en 1950 [Remontet 03]. Cette différence pourrait être liée à l'évolution des modes de vie décrits dans la partie 1.2.3, page 17 ou aux effets du dépistage généralisé.

1.2.2 Physiologie du cancer du sein

La fonction biologique du sein est de produire du lait pour nourrir un nouveau-né. Le sein est divisé en quinze à vingt compartiments, séparés par des tissus adipeux, qui sont chacun constitués de lobules et de canaux (voir figure 1.3). Les lobules produisent le lait en période d'allaitement et les canaux le transportent vers le mamelon [Luporsi 07].

Cancérogenèse : De manière générale, « le cancer est une tumeur liée à la

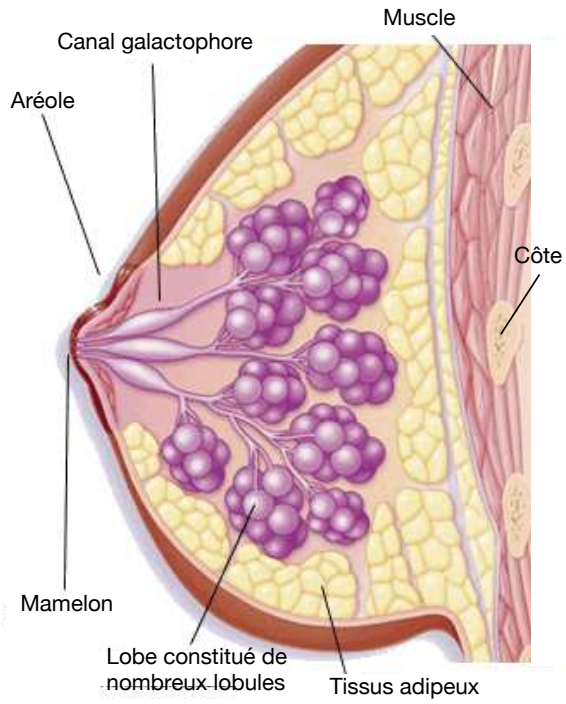

Figure 1.3 : Coupe anatomique du sein selon [LCC 12].

prolifération à la fois anarchique et indéfinie d'un clone cellulaire conduisant à la destruction du tissu originel, à l'extension locale, régionale et générale de la tumeur puis à la mort de l'individu en l'absence de traitement » [Kernbaum 08]. La croissance des cellules et des tissus est en effet assurée en différentes étapes régulées par un système complexe fondé sur l'information génétique. Cette information génétique est transmise entre les différentes générations de cellules et peut, malgré des processus de duplication et de réparation adaptés, être altérée. L'altération, éventuellement amplifiée par des agents cancérogènes, est facilitée par la préexistence de modifications germinales transmissibles. Elle conduit à la perte des mécanismes qui permettent le contrôle de la communication cellulaire indispensable au déclenchement de l'apoptose* et au contrôle de la prolifération des cellules.

On peut distinguer quatre phases de progression dans la cancérogenèse. Le processus débute par l'altération transmissible, et stable génétiquement, d'une cellule somatique* au cours de la phase d'initiation qui est irréversible. Le processus peut s'arrêter à ce stade d'hyperplasie* pendant de nombreuses années au cours desquelles les cellules initiées persistent dans l'organisme en latence. Le processus se poursuit par la phase de promotion, réversible, au cours de laquelle d'autres altérations conduisent au stade prolifératif aboutissant à l'apparition de lésions précancéreuses. La tumeur est alors installée, c'est le stade dysplasique. Enfin, la phase de progression, irréversible, correspond à l'acquisition de la malignité qui précède la

phase d'invasion à d'autres tissus. On parle alors de cancer invasif, sinon, on parle de cancer in situ.

Cancérogenèse mammaire : Concernant le cancer du sein, bien que ses mécanismes moléculaires de déclenchement ne soient pas complètement connus [Russo 00], on peut expliquer les différents types de cancers invasifs par le résultat d'une prolifération cellulaire incontrôlée ou d'un arrêt de l'apoptose dû à l'inactivation de gènes suppresseurs de tumeurs [Preston-Martin 90]. Le dysfonctionnement de ces processus peut être causé soit par des altérations transmissibles, soit par des carcinogènes environnementaux qu'ils soient biologiques, physiques ou chimiques. En parallèle, l'épithélium* mammaire sain serait atteint par des lésions prolifératives bénignes et atypiques, puis par un carcinome in situ, avant d'aboutir à une tumeur invasive [Allred 01, Burstein 04]. Cette hypothèse du continuum lésionnel faisant état d'une aggravation de la lésion de l'hyperplasie simple au carcinome invasif est toujours discutée [Dauplat 04].

Classification des tumeurs : Différentes classifications permettent de catégoriser les cancers du sein, notamment dans le but de préciser le pronostic de la maladie et le traitement envisagé. Les plus importantes sont décrites ci-dessous.

- Le grade histopronostique permet d'évaluer l'agressivité d'un cancer [Kapoor 05]. Le système le plus utilisé est le grade SBR (Scarff-Bloom-Richardson) qui permet de quantifier l'importance de la différenciation entre les tissus tubulaires et glandulaires, l'appréciation du degré d'anomalies nucléaires et le décompte des divisions mitotiques.
- La classification TNM, pour *Tumor* (tumeur), *Nodes* (ganglions) et *Metastasis* (métastases), permet d'évaluer le stade d'extension de la tumeur maligne en associant un chiffre mesurant l'évolution de chaque élément T, N et M caractérisant la tumeur.
- La présence de marqueurs biologiques et moléculaires. Par exemple, les récepteurs aux estrogènes et à la progestérone qui sont des hormones, dont les rôles sont de favoriser le développement des caractères sexuels féminins ainsi que de favoriser la nidation et la grossesse. La présence de la protéine membranaire HER2 est aussi une caractéristique importante. La présence ou l'absence de ces marqueurs constitue un facteur prédictif de la réponse aux thérapies qui visent à bloquer la prolifération des cellules [Kim 06].
- Le type histologique du cancer du sein qui permet de préciser le tissu à l'origine du cancer. L'atteinte des canaux galactophores représente la majorité des tumeurs malignes et est appelée carcinome au contraire des sarcomes qui touchent les tissus conjonctifs, lymphomes qui touchent les tissus lymphoïdes et métastases qui, regroupés, représentent moins de 2 % des cas de cancer du sein.

Ces différentes classifications applicables aux tumeurs du sein sont autant de critères qui permettent d'adapter le traitement pour obtenir de meilleures chances de gué-

rison dans le cadre de la médecine personnalisée. À terme, ce type de classification pourrait contribuer à affiner encore les politiques de prévention pour une prévention personnalisée.

1.2.3 Les facteurs de risque du cancer du sein

La prévention personnalisée, comprise au sens de l'assistance au praticien dans son évaluation du risque de maladie, nécessite de s'intéresser de près aux causes connues de la maladie. Notre objectif étant d'améliorer l'évaluation du risque de cancer de sein dans la population générale, il est utile de faire un état des lieux de la connaissance des facteurs à l'origine de ce cancer.

Outre l'âge, les facteurs de risque connus pour le cancer du sein se divisent en cinq grandes catégories, les voici par importance d'impact sur le niveau de risque. Les facteurs génétiques, et donc héréditaires, sont ceux qui augmentent le plus le risque de cancer du sein. Les facteurs hormonaux regroupent l'ensemble des risques liés à la vie reproductive des femmes. La densité mammaire dénote le nombre de cellules qui peuvent potentiellement être atteintes par un cancer. Les facteurs environnementaux englobent les habitudes alimentaires, l'activité physique, le statut tabagique et les caractéristiques anthropométriques, ils ont un faible impact sur le risque, mais ont l'avantage d'être modifiables. Enfin, à la différence des facteurs de risque qui sont au moins partiellement responsables d'une maladie, les marqueurs de risque, comme le fait d'avoir subi une biopsie mammaire, sont corrélés au risque sans pour autant que leur suppression soit envisageable ou entraîne une baisse effective du niveau risque.

Âge de la femme : En l'état actuel des connaissances, le premier facteur de risque pour le cancer du sein est l'âge de la femme (voir figure 1.4). Ce facteur n'appartient précisément à aucune des cinq catégories décrites précédemment. Il recouvre la part inexpliquée des causes de cancer du sein. À partir de 30 ans, l'incidence augmente de manière significative jusqu'à environ 50 ans, âge à partir duquel l'augmentation est plus modérée. Un palier est atteint entre 60 et 75 ans avant qu'une baisse du risque n'intervienne au-delà de 75 ans. Plusieurs hypothèses permettent d'expliquer cette baisse : l'effet cohorte pourrait en être à l'origine (voir page 14) notamment par l'intermédiaire d'une plus faible exposition aux traitements hormonaux de la ménopause, tout comme un dépistage moins fréquent ou encore un renouvellement cellulaire moins fréquent après 75 ans. Si l'étiologie* du cancer du sein n'est pas encore totalement connue, l'identification de différents facteurs, détaillés ci-après, permet d'expliquer en partie l'évolution de l'incidence de la maladie.

Facteur héréditaire : La génétique joue un grand rôle dans la compréhension de l'origine du cancer du sein. On peut différencier deux types de facteurs génétiques qui impactent le risque de cancer du sein et qui peuvent être utilisés dans des scores de risque (voir partie 2.2.1.1, page 31).

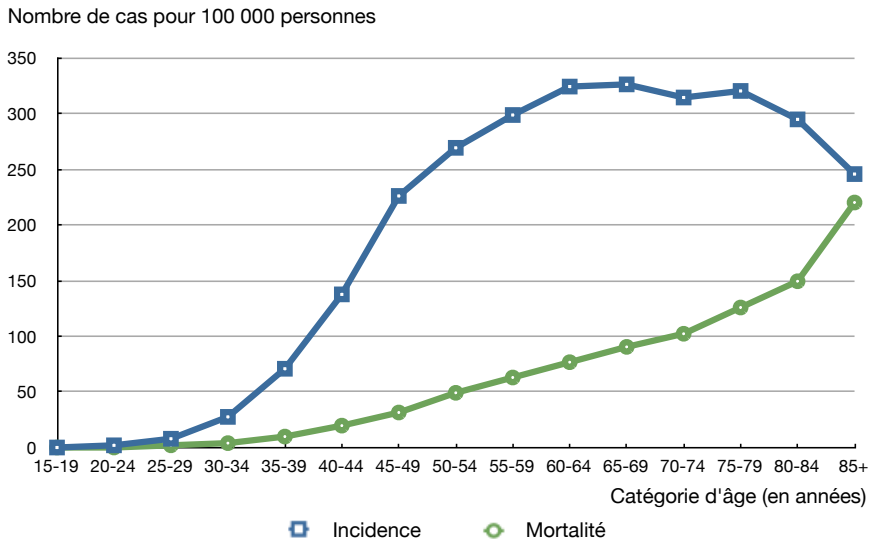

Figure 1.4 : Évolution des taux d'incidence et de mortalité pour l'année 2000 en France, pour 100 000 personnes-années pour le cancer selon [Trétarre 04].

D'une part, on parle de *prédisposition familiale* lorsque plusieurs cas de cancer du sein sont identifiés dans une même famille sans que la transmission entre les membres soit systématique. Elle est le résultat de l'effet cumulé de plusieurs facteurs de prédisposition qui peuvent être aggravés par les habitudes de vie de la femme. On parle de gène à faible pénétrance. À titre d'exemple, le risque d'une femme, dont une parente au premier degré (mère, sœur et fille) est atteinte, est multiplié par 1,9 [Pharoah 97] par rapport à une femme qui ne possède pas d'antécédents.

D'autre part, les *formes familiales* de cancer du sein sont liées à une mutation majeure de gènes précis. Ces formes familiales ne concernent qu'environ 5 % de la population, mais ces altérations entraînent l'apparition de cancers du sein avant l'âge de 40 ans et la multiplication du risque jusqu'à 1,8 [Ford 98]. Elles sont dues à la mutation des gènes BRCA 1 sur le chromosome 17 et BRCA 2 sur le chromosome 13 (BRCA signifie BReast CAncer). Ces gènes sont à l'origine de protéines dont le rôle est fortement lié à la réparation des doubles hélices d'ADN, le support de l'information génétique dans les cellules. La transmission de ces gènes est autosomique (sur les chromosomes non sexuels) et dominante (une seule version de l'allèle* est nécessaire pour l'expression du caractère). Plus globalement, d'autres mutations peuvent intervenir dans le risque de cancer du sein. Par exemple, en cas de mutation des gènes codant pour les protéines p53, et ATM [2], ce sont des fonctions liées à la division cellulaire et à l'apoptose, et respectivement à la réparation de l'ADN, qui seront perturbées.

2. Ataxia Telangiectasia Mutated, mutation qui provoque le syndrome d'ataxie télangiectasie.

Facteur hormonal : Le cancer du sein étant un cancer hormono-dépendant [Key 88], la durée de l'imprégnation hormonale de la femme est donc un facteur de risque. L'imprégnation hormonale est l'exposition aux hormones endogènes* et exogènes*.

Deux facteurs en particulier permettent de caractériser l'exposition aux hormones endogènes, notamment produites par les ovaires. L'âge aux premières règles marque le début de l'exposition hormonale. Si ce début d'exposition est précoce, le risque augmente [Key 01]. L'exposition se poursuivra jusqu'à la ménopause. Si celle-ci est tardive, la durée de l'exposition augmente et le risque de cancer du sein également. Selon [Group 97], le risque augmenterait de 3 % pour chaque année supplémentaire d'exposition aux hormones à partir de l'âge de la ménopause, par exemple à cause de la prise de traitement substitutif de la ménopause. En conséquence, les facteurs venant modifier le déroulement classique des cycles menstruels peuvent influer sur le risque. Par exemple, l'âge à la première grossesse influe [Rosner 94] et chaque grossesse influe sur le risque en deux temps : une légère augmentation du risque dans les 10 années qui suivent la fin de la grossesse puis une diminution du risque à long terme [Li 00]. Un nombre d'enfants [Clavel-Chapelon 02] élevé et une durée d'allaitement élevée modifient quant à eux le risque à la baisse [Group 02].

L'exposition aux hormones exogènes est principalement due aux contraceptifs oraux et aux traitements de la ménopause. Chez les femmes qui utilisent couramment des contraceptifs oraux, le risque de cancer du sein est augmenté jusqu'à 25 % environ, mais cet accroissement disparaît après 10 années d'arrêt d'utilisation [Group 96]. Le type de molécule et la durée d'utilisation ne semblent pas avoir d'effet sur le niveau de risque. Pour les traitements hormonaux de la ménopause (THM), qui permettent de pallier les troubles climatériques* de la ménopause, les études vont dans le sens d'une augmentation du risque proportionnelle à la durée d'utilisation, lors de la prise d'un estro-progestatif (estrogène combiné à un progestatif de synthèse) [Rossouw 02, Beral 03, Fournier 05, Fournier 08] avec un arrêt de la période à risque environ cinq ans après l'arrêt du traitement.

Densité mammaire : La densité mammaire constitue un indice de mesure de la quantité de tissus adipeux et de tissus épithéliaux présents dans le sein. Elle est détectable par radiographie mammaire. Plusieurs études ont montré que la densité mammaire influait sur le risque de cancer du sein [McCormack 06, Varghese 12] probablement parce qu'une forte densité mammaire dénote une forte présence de cellules mammaires, celles-ci étant le support des lésions cancéreuses dans le cancer du sein. La densité mammaire fluctue au cours de la vie reproductive et en fonction de paramètres comme la consommation d'alcool ou l'indice de masse corporelle (rapport du poids (kg) sur la taille (en mètre) au carré). Il est à noter que les variations de densité mammaire entre individus s'expliquent en grande partie par des causes génétiques [Boyd 05].

Nutrition et habitudes de vie : Alcool mis à part [Cottet 09], aucun aliment

n'est précisément mis en cause par les études sur le risque de cancer du sein, l'hypothèse d'un effet de la nutrition sur le niveau de risque est étudiée sous différents angles. Par exemple, le fait que les femmes émigrant d'un pays à faible incidence de cancer du sein (et à niveau de vie comparable) voient leur descendance rattraper le niveau de celui du pays d'accueil en deux à trois générations laisse penser que l'alimentation pourrait être à l'origine de ce phénomène [Buell 73, Stanford 95]. En l'état actuel des connaissances, l'apport énergétique, les graisses alimentaires de manière générale et les fibres ne semblent pas favoriser le cancer du sein. En revanche, les apports en acide gras trans (issus de la transformation dans l'industrie agroalimentaire) augmente le risque [Chajès 08]. A contrario, la consommation de légumes [Gandini 00] et les apports en acide gras polyinsaturés [Chajès 08] réduiraient le risque de cancer du sein.

Avant la ménopause, l'excès de poids n'entraîne pas d'augmentation du risque du cancer du sein, notamment parce que le tissu adipeux* agit comme un site de stockage des hormones, diminuant ainsi l'exposition des tissus mammaires aux hormones (voir l'imprégnation hormonale page 19). En revanche, après la ménopause, le surpoids et l'obésité sont des facteurs de risque avérés de cancer du sein [Key 01], notamment parce que les tissus adipeux sont une source d'estrogènes dans le corps.

Il semblerait que l'activité physique régulière soit un facteur de protection vis-à-vis du cancer du sein [Key 01, Friedenreich 08]. Un bénéfice maximal semble tiré d'une activité physique intense, la production d'estrogènes étant réduite dans ce cas là [Tehard 06, Lynch 11].

Bien que la fumée de cigarette soit une source importante de substances cancérigènes, pendant longtemps, le statut tabagique n'a pas été considéré comme un facteur de risque établi pour le cancer du sein. Certaines études montraient un risque réduit pour les fumeuses à cause d'un effet anti-estrogène [MacMahon 80] et d'autres un risque augmenté [Reynolds 04]. Cependant une méta-analyse récente a de nouveau mis en avant une augmentation du risque associé à la consommation de tabac [Collishaw 09].

Marqueurs de risque : D'autres éléments permettent de mesurer le risque sans pour autant qu'ils soient des facteurs à l'origine du cancer du sein. C'est le cas du fait d'avoir subi une biopsie mammaire. En effet, la biopsie n'influe pas sur le risque de cancer du sein, mais permet de marquer, de manière plus ou moins précise selon les populations, un risque plus élevé de cancer.

Les maladies bénignes du sein sont également un marqueur de risque au sens où certains types de lésions (non-prolifératives ou prolifératives sans atypie*) sont associées à une légère augmentation du risque de cancer du sein tandis que les lésions prolifératives avec atypies sont associées à une multiplication du risque par quatre [Dupont 87]. Comme pour les biopsies, il ne faut pas confondre association avec le cancer et cause de cancer.

Enfin, certaines maladies peuvent être associées à une modulation du risque, ce serait par exemple le cas du diabète qui serait associé à une hausse du risque de

cancer du sein [Xue 07, Fagherazzi 11].

Bien que le cancer du sein soit une maladie très étudiée pour laquelle on connaît à la fois les grandes étapes du processus de cancérogenèse mammaire et l'effet de nombreux facteurs de risque, il reste encore des inconnues notamment du point de vue des processus précis de cancérogenèse et des facteurs de risque. L'âge reste un élément de prédiction majeur du risque de cancer du sein, mais ne peut être utilisé seul pour estimer le risque.

La prévention du cancer du sein peut être améliorée au travers d'une amélioration du dépistage par un meilleur ciblage. Cette personnalisation de la prévention entre pleinement dans les objectifs des acteurs du monde de la santé dont la Fondation ARC qui finance une clinique du risque qui permettra la mise en œuvre concrète de ces principes de personnalisation du traitement et de la prévention.

2 Les scores de risque et la mesure de leur performance

La création et l'utilisation de scores de risque dans le domaine de la santé sont relativement récentes dans l'histoire des sciences, la cohorte de patients ayant servi à la création du premier score de risque construit en population générale date en effet de 1948. Dans ce chapitre, nous expliquons l'origine des scores de risque dans le domaine de la santé, en évoquant le cas particulier des maladies cardio-vasculaires qui ont bénéficié des premiers outils d'estimation diffusés à l'échelle mondiale. Nous présentons également les scores de risque majeurs pour le cancer du sein qu'ils soient familiaux ou génétiques. Enfin, nous détaillons les différentes mesures que nous utiliserons pour évaluer la performance des scores de risque créés, que ce soit en termes de discrimination ou de calibration.

2.1 Origine des scores de risque en santé

La généralisation de la production de scores de risque pour diverses maladies a été précédée d'une phase pendant laquelle les scores de risque étaient essentiellement construits pour le risque cardio-vasculaire. Dans cette partie, nous abordons les premiers scores en médecine, les données à l'origine des premiers scores de risque et deux modèles statistiques utilisés pour construire ces scores de risque.

2.1.1 Les scores descriptifs

Il existe en médecine de très nombreux scores, parfois appelés échelles, mesures ou tests. Historiquement, les premiers d'entre eux ont servi à décrire l'état, et éventuellement la progression, de la maladie, de la douleur ou de l'état d'esprit d'un patient. Le résultat pouvant servir à aider ou à justifier la décision d'un médecin, par rapport à un traitement par exemple, on parlera alors plus précisément de score diagnostic. Par exemple, l'indice de masse corporelle (IMC) permet d'évaluer rapidement, par le simple rapport masse (en kilogramme) sur taille (en mètre) au carré, l'éventuel surpoids d'une personne grâce à une échelle standardisée par l'OMS [WHO 00]. Si le rapport est inférieur à 18,5 $\mathrm{kg{\cdot}m^{-2}}$, on parle de maigreur puis de dénutrition. Si le rapport est supérieur à 25 $\mathrm{kg{\cdot}m^{-2}}$, on parle de surpoids puis d'obésité modérée, sévère et morbide.

Pour faciliter l'utilisation des scores, les auteurs ont pu choisir d'utiliser un système de points attribués en fonction de réponses fournies par le patient ou le médecin lui-même. La somme des points permet au médecin de placer le patient, pour un critère donné, sur une échelle globale afin d'améliorer sa prise en charge. Par exemple, l'échelle de [Hamilton 60] permet, dans sa version originale, de mesurer l'état dépressif d'un sujet grâce à 17 questions à choix multiples en évaluant les grandes familles de facteurs de risque identifiés par l'auteur que sont l'humeur, le sommeil, l'anxiété ou l'alimentation. Les scores peuvent aussi être utiles en situation d'urgence pour permettre à un médecin d'évaluer le plus rapidement et le plus objectivement possible la situation d'un patient. Par exemple, l'échelle de Glasgow, formalisée par [Teasdale 74], permet d'évaluer l'état neurologique d'une personne en fonction de critères liés à sa réponse verbale, oculaire ou motrice. Plus la réponse est prononcée, plus le nombre de points attribué est élevé et moins l'état du patient est critique. Utilisé sur la durée, le score permet notamment de juger de l'évolution de l'état du malade et de l'impact éventuel d'un acte chirurgical.

Le champ d'application des scores est vaste : certains permettent en effet d'évaluer des éléments liés à la qualité de vie comme le stress [Cohen 83], la dépendance à la nicotine [Fagerström 78], la motivation à l'arrêt du tabac [Lagrue 02], la qualité de l'alimentation [Vercambre 09], la dépendance à l'alcool [Saunders 93] ou encore le fonctionnement de capacités cognitives avec le test de [Folstein 75]. Ils permettent de décrire l'état d'un patient au moment de l'utilisation du score, et non le risque futur.

2.1.2 Des premiers scores prédictifs complexes

Si les premiers scores présentés ci-dessus sont essentiellement descriptifs, le suivi d'individus au sein de cohortes prospectives* a été l'occasion pour les épidémiologistes de mettre au point des scores de risque par maladie. Il s'agit donc de calculer une probabilité de survenue d'une maladie dans le futur au lieu de décrire l'état actuel d'un individu. L'exercice est rendu d'autant plus difficile que la maladie a une prévalence faible, c'est-à-dire qu'une part relativement faible de la population en est atteinte.

Nous présentons d'abord les données qui ont permis l'émergence des premiers scores de risque puis deux modèles statistiques (utilisés pour construire de nombreux scores de risque) au travers de l'exemple du risque cardio-vasculaire tout en mettant en avant la complexité intrinsèque des modèles et la difficulté à faire évoluer les scores de risque produits.

La cohorte de Framingham : Avec la « Nurses Health Study » et la « British Medical Doctors Study », la cohorte de Framingham est une des plus anciennes cohortes. Elle a permis l'émergence du premier score de risque utilisé à l'échelle mondiale. À partir de 1948, elle a permis de suivre, dans sa première version, 5 209 personnes âgées de 30 à 62 ans, habitant la ville de Framingham aux États-Unis et

choisies pour leur représentativité de la population américaine de l'époque. Depuis cette date, les membres de la cohorte (et des autres cohortes qui ont été constituées sur ce modèle dans la ville) font régulièrement l'objet d'examens médicaux qui sont l'occasion d'un recueil de données : des paramètres permettant de caractériser des profils sont recueillis et les occurrences de maladies sont relevées. La cohorte est donc de type prospective* puisque la mesure des expositions est effectuée avant la survenue des maladies pour éviter les biais de mémoire différentiels, à la différence des études de cohortes rétrospectives* pour lesquelles il faut reconstituer les éléments du suivi d'exposition après la survenue des maladies.

À son lancement, par ce qui est devenu l'institut national du cœur, des poumons et du sang aux États-Unis, la cohorte de Framingham avait pour objectif principal l'étude du risque cardio-vasculaire. Les éléments recueillis étaient, par hypothèse, liés à cette maladie et ont permis d'établir des associations entre le risque cardio-vasculaire et le niveau de cholestérol, la pression artérielle et les irrégularités d'électrocardiogramme [Kannel 61], puis l'influence du surpoids et de l'activité physique [Kannel 67] et la survenue de la ménopause [Kannel 76a]. Après la mise en évidence de plusieurs facteurs de risque, les investigateurs ont choisi de construire un score regroupant les différents facteurs de risque connus pour les maladies cardio-vasculaires afin d'estimer le risque global [Kannel 76b]. La modélisation du risque global a été mise à jour plusieurs fois et une révision majeure a été produite par [Anderson 91].

Cette modélisation du risque fait partie d'une branche des statistiques appelée *analyse de survie*. Pour un événement donné qui n'est pas forcément la mort (cardiopathie coronarienne dans l'exemple du score de Framingham), l'objectif est de prendre en compte le temps avant occurrence de cet événement. On exprime donc la survie de manière qualitative, par l'occurrence de l'événement ou non, et de manière quantitative, par la durée avant l'événement. La fonction de survie S, qui correspond à la probabilité que l'événement intervienne après le temps t, peut s'écrire : $S(t) = P(T \geq t)$ pour $t \geq 0$, où t est la variable temps et T une variable aléatoire symbolisant le moment où se produit l'événement. Cette fonction de survie peut être fonction d'un vecteur d'attributs $Z_i = (Z_1, Z_2, \ldots, Z_n)$ qui décrivent l'individu.

Différents modèles statistiques permettent de réaliser ce type d'analyse de survie. Détaillons l'utilisation d'une modélisation par temps de défaillance, puis d'une modélisation par modèle à risques proportionnels dont l'utilisation est fréquente.

Un modèle de temps de défaillance : Basé sur des données plus nombreuses pour ce qui est des sujets et des événements cardio-vasculaires, plus précises en termes d'analyses biologiques et plus récentes que celles utilisées pour les précédents travaux de construction de score sur la cohorte originelle de Framingham, le score de risque d'[Anderson 91] est construit sur un modèle de temps de défaillance accéléré. Il s'agit d'un modèle de survie paramétrique ce qui signifie qu'une loi de distribution doit être spécifiée pour la variable T, ici une loi de Weibull qui, selon les auteurs, est souvent utilisée pour les analyses portant sur le temps avant événement. La

probabilité p que l'événement se produise pendant un temps t s'écrit :

$$p(t) = 1 - \exp(-\exp(u)) \text{ avec } u = \frac{\log_e(t) - \mu}{\sigma}, \tag{2.1}$$

où le logarithme utilisé est en base exponentielle (logarithme népérien), le paramètre d'échelle [1] σ peut s'exprimer sous la forme $\log_e(\sigma) = \theta_0 + \theta_1 \cdot \mu$ pour une meilleure adéquation du modèle. Dans les deux cas, le paramètre de position [2] μ correspond à la combinaison linéaire des Z_i et des β_i, ce dernier étant les coefficients de régression des Z_i, tel que

$$\mu = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_n Z_n \,. \tag{2.2}$$

Après estimation des paramètres par la méthode du maximum de vraisemblance, les auteurs proposent quatre équations pour évaluer le risque de survenue de cardiopathie coronarienne, en fonction du sexe de l'individu concerné et du type de pression artérielle retenu. Pour une femme, avec une mesure de pression artérielle de type systolique et avec « **t** » l'échéance du risque en années, « **pression** » la pression artérielle systolique en mmHg, « **choltot** » le niveau de cholestérol total en mg·dl^{-1}, « **cholhdl** » le niveau de cholestérol de type HDL (pour « High Density Lipoproteins »), « **tabac** » le statut tabagique codé en 0 ou 1 et « **âge** », l'âge de la personne, l'équation proposée par les auteurs est la suivante :

$$p(t) = 1 - \exp\left(-\exp\left(\frac{\log_e(\mathbf{t}) - (m + 4,418\,1)}{\exp(-0,315\,5 - 0,278\,4\,m)}\right)\right), \tag{2.3}$$

où m vaut, dans le cas d'une femme :

$$m = a - 5,854\,9 + 1,851\,5[\log_e(\mathbf{\hat{a}ge}/74)]^2\,, \tag{2.4}$$

avec a valant :

$$\begin{aligned} a = 11,112\,2 - 0,911\,9\log_e(\mathbf{pression}) - 0,276\,7\,\mathbf{tabac} \\ -0,718\,1\log_e(\mathbf{choltot}/\mathbf{cholhdl})\,. \end{aligned} \tag{2.5}$$

Pour une femme dont on veut calculer le risque à 10 ans, qui est âgée de 55 ans avec une pression artérielle systolique de 130 mmHg, un niveau de cholestérol total mesuré à 240 mg·dl^{-1}, un niveau de cholestérol HDL mesuré à 45 mg·dl^{-1} et qui fume : la probabilité de survenue d'un événement coronarien est $p = 0,135$ pour le profil évalué, c'est-à-dire 13,5 % de risque d'être victime d'un tel événement à 10 ans. La traduction de ces équations d'estimation du risque en deux scores à points (voir page 24), selon l'échéance de calcul du risque à 5 ou 10 ans, permet grâce à une table de correspondance point vers probabilité, d'obtenir une évaluation approchée

1. Un paramètre d'échelle régit l'aplatissement d'une famille paramétrique de lois de probabilités.

2. Un paramètre de position régit la tendance centrale (moyenne, mode, médiane) d'une densité de probabilité.

du risque. Malgré cette disposition, la compréhension du modèle statistique reste difficile pour un utilisateur n'ayant pas de formation statistique. De plus, dans le cas où de nouvelles données seraient mises à disposition au cours du cycle de vie du score ou si un ou plusieurs facteurs de risque n'étaient pas disponibles dans un contexte d'utilisation précis, l'équation ne pourrait être facilement mise à jour, car elle n'est pas évolutive, la valeur des paramètres du modèle ayant été optimisée pour une meilleure adéquation du modèle. La mise à jour du modèle pour une de ces raisons entraînerait la modification de tous les coefficients de la régression, ce qui conduirait à une équation différente de celle présentée, sans qu'il soit possible de mettre en évidence une logique apparente dans les modifications. Une mise à jour serait en outre complexe à mettre en œuvre puisqu'elle nécessiterait de refaire le processus de construction.

Un modèle à risques proportionnels : D'autres scores de risque ont également été construits sur les données des cohortes de Framingham, c'est le cas du score de [D'Agostino 08]. Son objectif est d'évaluer le risque cardio-vasculaire au sens large puisqu'il ne s'agit plus seulement de prendre en compte les cardiopathies coronariennes, mais également les maladies cérébro-vasculaires, les artéropathies des membres inférieurs et les insuffisances cardiaques sévères, dont l'infarctus du myocarde.

Un modèle de survie, semi-paramétrique cette fois, est utilisé pour modéliser le risque : le modèle à risques proportionnels est souvent abrégé par *régression de Cox* ou *modèle de Cox*, du nom du statisticien anglais David Cox qui l'a popularisé. Ce modèle permet d'intégrer simultanément l'effet de plusieurs facteurs pour augmenter la puissance du modèle et éliminer les facteurs de confusion* connus ou suspectés [Timsit 05]. Il est semi-paramétrique car pour expliquer la survenue d'un événement, il n'est pas nécessaire de faire d'hypothèse sur la forme de la fonction de survie. Tout comme la régression linéaire ou logistique, il s'agit d'une méthode de régression multivariée dont l'objectif est de modéliser la variable T symbolisant le moment où se produit l'événement. La régression de Cox modélise le risque instantané, en fonction des attributs $Z_i = (Z_1, Z_2, \ldots, Z_n)$ qui décrivent l'individu, sous la forme d'une multiplication :

$$h(t) = h_0(t)\, \exp(\beta' Z_i)\,, \tag{2.6}$$

où $h_0(t)$ est une fonction de risque de base commune à tous les individus et où $\exp(\beta' Z_i)$ représente une fonction de régression dans laquelle β' représente les coefficients de régression inconnus. Deux hypothèses sont à vérifier pour utiliser ce modèle : d'une part qu'il existe une relation log-linéaire entre les attributs et la fonction de risque et d'autre part, que le rapport des fonctions de risque instantané de deux sujets ne dépend pas du temps (proportionnalité des risques).

La formule générale pour le calcul de la probabilité d'un événement est :

$$p(t) = 1 - h_0(t)\, \exp\left(\sum_{i=1}^{n} \beta_i Z_i - \sum_{i=1}^{n} \beta_i \bar{Z}_i\right)\,, \tag{2.7}$$

où $h_0(t)$ est toujours la fonction de risque de base, β_i le coefficient de régression estimé, Z_i la valeur du $i^{\text{ème}}$ facteur de risque, $\bar{Z}_i$ la moyenne correspondante et n le nombre de facteurs de risque.

En utilisant ce modèle, [D'Agostino 08] construit un score de risque basé sur 8 491 individus. Le score est construit à partir des mêmes facteurs de risque identiques à ceux du modèle d'[Anderson 91] (page 26) auxquels ont été ajoutés la présence ou l'absence d'un traitement pour la tension artérielle symbolisée dans l'équation par « **ttt** » et le statut diabétique « **diabète** ». Après estimation des paramètres, les auteurs obtiennent l'équation suivante pour la partie $\sum_{i=1}^{n} \beta_i Z_i$ de l'équation 2.7 :

$$\begin{aligned}\sum_{i=1}^{n} \beta_i Z_i = {} & 2,328\,88 \log_e(\textbf{âge}) + 1,209\,04 \log_e(\textbf{choltot}) \\ & -0,708\,33 \log_e(\textbf{cholhdl}) + 2,761\,57 \log_e(\textbf{pression}) \\ & +2,822\,63\,\textbf{ttt} + 0,528\,73\,\textbf{tabac} + 0,691\,54\,\textbf{diabète}\,. \end{aligned} \tag{2.8}$$

Pour l'exemple, on cherche à estimer le risque à 10 ans d'un individu au profil identique que celui utilisé précédemment, c'est-à-dire une femme de 55 ans avec une pression artérielle systolique de 130 mmHg, un niveau de total cholestérol mesuré à 240 mg·dl^{-1}, un taux de cholestérol de type HDL à 45 mg·dl^{-1} et qui fume. Des paramètres auxquels on ajoute l'absence de diabète et l'absence de traitement pour la tension artérielle. L'équation 2.9 permet de décomposer le calcul à l'aide de $\sum_{i=1}^{n} \beta_i Z_i$ détaillé dans l'équation 2.8, de $\sum_{i=1}^{n} \beta_i \bar{Z}_i$ dont les valeurs sont fixées pour la population d'analyse et du risque global :

$$\begin{aligned} p &= 1 - h_0(t)\, \exp\left(\sum_{i=1}^{n} \beta_i Z_i - \sum_{i=1}^{n} \beta_i \bar{Z}_i\right) \\ &= 1 - 0,950\,12\, \exp(27,233\,3 - 26,193\,1) \\ &= 0,134\,8 = 13,5\,\%. \end{aligned} \tag{2.9}$$

Cette femme a donc 13,5 % de risque d'être victime d'un événement cardio-vasculaire au sens large à un horizon de 10 ans.

Comme pour le modèle d'Anderson *et al.*, un système de score à points est proposé. Il permet de calculer de manière plus simple, mais approchée, le résultat fourni par l'utilisation de l'équation grâce à une addition de points effectuée en fonction des réponses fournies. Malgré ce système, tout comme le modèle de temps de défaillance, le modèle de proportionnalité des risques met en jeu l'estimation de paramètres d'un modèle statistique de survie qui complique l'accès à la compréhension du score. Comme le modèle d'Anderson, ce score de [D'Agostino 08] apparaît comme un modèle opaque à tout utilisateur qui n'aurait pas de connaissances en statistique. De même, le changement de la population, des paramètres à disposition pour estimer le risque ou du contexte de son utilisation nécessiterait une reconstruction complète du score qui mènerait à la production d'une équation différente de celle proposée.

En effet, l'utilisation des scores de risque devrait être limitée à des populations comparables à la population qui sert à générer le score. Le changement de population est donc nécessaire pour adapter le score à d'autres pays ou à d'autres sous-populations.

2.1.3 Généralisation des scores de risque

La validation des scores conçus pour le risque cardio-vasculaire sur d'autres populations a nécessité de mener des études dans d'autres pays ou sur d'autres sous-populations. De même, la création de scores de risque pour d'autres maladies a nécessité l'application des modèles, ceux présentés et d'autres, sur d'autres bases de populations.

Validation sur d'autres populations : La construction de scores sur la base de données de cohorte, aussi représentatives qu'elles puissent être des populations des pays concernés, ne préjuge pas de l'exportabilité à d'autres populations. Outre la validation interne du score qui permet de vérifier qu'il est performant sur l'échantillon de population qui a servi à sa construction, il faut également mesurer sa validité externe. L'application du score à un autre échantillon de la même population permet de mesurer sa reproductibilité tandis que l'application du score à une population différente permet de vérifier sa généralisation (ou transportabilité).

Historiquement, les scores de risque cardio-vasculaire étant les plus anciens, ils ont été les premiers à être conçus à travers le monde. Mais, le rapport de l' [ANAÉS 04] met en avant l'hétérogénéité des populations, des méthodes de prédiction, des méthodes de mesure des attributs utilisés et des définitions de l'événement étudié. Des études transversales ont en effet été réalisées pour tester les scores de risque sur d'autres populations que celles ayant servi à construire les scores. Toujours en prenant l'exemple du risque cardio-vasculaire, l'étude de [Dréau 01] a permis d'analyser le risque prédit par 32 modèles de risque sur une population de sujets avec une pression artérielle plus élevée que la moyenne. L'étude de [Laurier 94] s'est, quant à elle, attachée à adapter le modèle d'[Anderson 91], décrit page 25, à une population française tandis que celle de [Haq 99] a permis de tester la validité de trois modèles de calcul du risque dont celui d'[Anderson 91]. Il ressort de ces études que la concordance* entre les différents scores de risque est satisfaisante pour les patients ayant un risque élevé ou faible, mais médiocre chez les sujets à risque moyen [Dréau 01]. Selon [Haq 99], le modèle d'Anderson *et al.*, construit sur la cohorte de Framingham, surestime le risque d'incident coronarien prévu par celui de l'étude BRHS (British Regional Heart Study). Concernant la population française, le modèle d'[Anderson 91] surestime le risque dans 70 % des cas, le sous-estime dans 1 % et concorde dans 29 % des cas. Une légère recalibration, par la modification du paramètre constant de l'équation de régression 2.5, page 26, de 11,11 22 à 11,44, permet cependant d'améliorer la performance du modèle en réduisant la surestimation à 12 % des cas et en augmentant la concordance à 80 % des cas selon [Laurier 94].

Une fois publié, un modèle n'est donc valable que pour la population d'origine sur laquelle il a été conçu. Pour être utilisé sur d'autres populations, dans d'autres pays par exemple, il doit non seulement passer par une étape de validation interne, mais aussi par une étape d'adaptation des modèles ou de recalibration pour chaque population pour lesquelles il est prévu de l'utiliser. Quand l'adaptation du score de risque n'est pas envisageable, il est toujours possible de construire de nouveaux modèles de risque comme le propose [Vergnaud 08].

Des scores de risque pour les grandes maladies : Si le risque cardio-vasculaire a été le premier à bénéficier de la construction de modèles de prédiction du risque et des études de validation dans d'autres populations, le risque de plusieurs maladies à forte prévalence dans la population a également été modélisé dans de nombreuses études. Le risque de déclin cognitif a, par exemple, été modélisé par [Kivipelto 06] grâce à une régression linéaire dont les coefficients ont permis la construction d'un score à points pour faciliter la détermination du risque de démence sur 20 ans. Les cancers en général, et ceux à forte prévalence en particulier, ont également fait l'objet de scores de risque : par exemple, [Driver 07] a conçu un score de risque pour le cancer colorectal sur une population d'hommes grâce à une régression linéaire et [Bach 03] a modélisé le risque de cancer du poumon pour les fumeurs à des fins de ciblage du dépistage. Le cancer du sein a, lui aussi, fait l'objet de nombreux scores de risque qui méritent d'être analysés plus en détail.

2.2 Scores de risque pour le cancer du sein

Puisque nous avons choisi de présenter une application de nos travaux au cancer du sein, il convient de s'intéresser plus particulièrement aux forces et aux faiblesses des scores de risque existant dans ce domaine avant d'expliciter nos objectifs sur les scores de risque que nous souhaitons construire et le processus qui permet d'y parvenir.

Dans un premier temps, nous montrons l'apport des scores des risques pour le cancer du sein issus de l'épidémiologie avant de constater, dans un second temps, que dans le domaine de la fouille de données, il n'existe pas de travaux traitant de la conception d'un score de risque en population générale pour le cancer du sein.

2.2.1 Scores issus de l'épidémiologie

Si plusieurs types de scores de risque pour le cancer du sein en prévention sélective sont apparus en même temps, les premiers scores utilisés pour déterminer le risque de cancer du sein ont essentiellement été basés sur des modèles familiaux, qu'ils soient empiriques ou génétiques.

2.2.1.1 Modèles familiaux : empiriques ou génétiques

Parmi les scores familiaux, la première catégorie regroupe les scores dits empiriques. Il s'agit des modèles qui ont pour objectif d'estimer le risque d'être porteur d'une mutation génétique augmentant très fortement le risque de cancer du sein. Ces modèles empiriques ne font pas d'hypothèse explicite sur le risque génétique, tant sur le plan de la pénétrance* du gène à risque dans la population, que sur le mode de transmission ou de la fréquence de mutation. C'est le cas du modèle de [Couch 97] ou du modèle de [Shattuck-Eidens 97]. Ce dernier permet par exemple de calculer le risque de cancer du sein pour les femmes en fonction du fait qu'elles aient été atteintes par un cancer de l'ovaire ou non : ce modèle a été construit grâce à un échantillon de 798 personnes, en utilisant un modèle de régression logistique dans lequel la probabilité p que l'événement « Être porteur d'une mutation délétère du gène BRCA1 » soit réalisé, est exprimée comme suit :

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_n Z_n\,, \tag{2.10}$$

où les paramètres Z_i correspondent aux facteurs de risque et les β_i aux coefficients de la régression logistique, ce qui permet aux auteurs de proposer l'équation de risque :

$$p = \frac{\exp(L)}{1+\exp(L)}\,, \tag{2.11}$$

avec :

$$L = -0,080\mathbf{a}+1,141\mathbf{b}+0,0\mathbf{c}+1,29\mathbf{d}+2,08\mathbf{e}+3,39\mathbf{f}+1,68\mathbf{g}+0,31\mathbf{h}+1,06\mathbf{i}+1,68\mathbf{j}\,,$$

dans laquelle :

- **a** représente l'âge au diagnostic du cancer du sein (ou cancer de l'ovaire),
- **b** l'origine ethnique,
- **c**, **d**, **e** et **f** la latéralité du cancer en lien avec la présence ou l'absence d'un cancer de l'ovaire,
- **g** la présence d'un cancer de l'ovaire,
- **h** le nombre d'antécédents familiaux de cancer du sein sans cancer de l'ovaire,
- **i** le nombre d'antécédents familiaux avec un cancer de l'ovaire sans cancer du sein,
- **j** le nombre d'antécédents familiaux avec les deux cancers.

La conception de ce score de risque illustre la volonté de prendre en compte avant tout la composante génétique du cancer du sein dans l'estimation du risque. Ce type de modèle, et en particulier ce modèle-ci n'intègre pas de facteurs de risque individuels.

Parmi les scores familiaux, une seconde catégorie regroupe les scores de risque dits génétiques. Il s'agit des scores pour lesquels les auteurs ont fait des hypothèses explicites sur les gènes, et leurs allèles*, impliqués dans le cancer du sein. C'est le cas du modèle de [Claus 91] qui, avant même la découverte de gènes dont on sait désormais qu'une mutation peut favoriser le cancer du sein, permettait d'expliquer la répartition des cancers du sein dans une famille par la transmission autosomale d'un gène à forte pénétrance, mais dont les mutations dans la population sont rares. L'analyse avait été menée lors d'une étude cas-témoin* en utilisant les mères et les sœurs de la cohorte CASH (pour « Cancer and Steroid Hormone Study ») grâce à une analyse de ségrégation dont l'objectif est de déterminer quel mode de transmission explique le mieux les distributions familiales observées. Ce modèle est disponible sous la forme de tableaux de risque faciles à lire dans lesquels on voit rapidement l'évolution du risque (voir tableau 2.1). En revanche, peu de facteurs peuvent être intégrés dans un tableau à double entrée et aucun facteur de risque individuel autre que l'âge de la patiente n'est utilisé. En outre, le modèle repose sur des hypothèses non vérifiées de la transmission des phénotypes avec une tendance à la sous-estimation des cancers du sein [De Pauw 09]. Le modèle BODICEA (pour « Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm »), dont les paramètres sont également dérivés d'une analyse de ségrégation, permet de calculer le risque de cancer, mais uniquement pour une famille dans laquelle on compte déjà des gènes BRCA mutés. Il n'intègre pas de facteurs de risque individuels, mais permet de prendre en compte l'origine polygénique* du cancer du sein, en tenant compte de l'effet additif de chaque gène muté sur le cancer du sein.

Un des plus utilisé est le modèle de Tyrer-Cuzick (aussi appelé IBIS) qui intègre, en plus du risque lié au génotype de la personne, des facteurs de risque individuels [Tyrer 04]. Au contraire des précédents modèles évoqués, celui-ci n'est pas basé sur la modélisation d'un échantillon de femmes, mais sur une combinaison de risques relatifs. Plusieurs améliorations de ce modèle ont été proposées, notamment celle de

Tableau 2.1 : Tableau des risques en fonction de l'âge du sujet et de l'âge du cancer du sein chez le parent au premier degré, d'après le modèle de [Claus 91].

Âge de	Âge du cancer chez le parent au 1er degré					
la femme	20-29	30-39	40-49	50-59	60-69	70-79
29	0,007	0,005	0,003	0,002	0,002	0,001
39	0,025	0,017	0,012	0,008	0,006	0,005
49	0,062	0,044	0,032	0,023	0,018	0,015
59	0,116	0,086	0,064	0,049	0,040	0,035
69	0,171	0,130	0,101	0,082	0,070	0,062
79	0,211	0,165	0,132	0,110	0,096	0,088

[Santen 07] qui intègre la prise en compte de la densité mammaire, de la taille et du rapport tour de taille sur tour de hanche.

2.2.1.2 Modèles basés sur des facteurs individuels

Les modèles présentés jusqu'à présent permettent d'estimer, soit le risque de porter une mutation génétique conduisant à une augmentation du risque de cancer du sein, soit le risque global de cancer du sein. Cependant, la plupart du temps, ces modèles ne font pas faire appel à des facteurs de risque individuels. Nous présentons dans cette partie deux modèles importants, par leur diffusion et leur utilisation, qui utilisent des facteurs individuels pour évaluer le niveau de risque de cancer du sein.

Modèle de Gail : Le modèle de [Gail 89], bien que faisant appel à la composante génétique du cancer du sein par l'intermédiaire de l'utilisation de nombre d'antécédents de cancer du sein, permet d'estimer un risque global grâce à l'intégration de facteurs individuels. Construit par Gail *et al.* en 1989, le modèle est basé sur une population issue d'une étude cas-témoin nichée dans la cohorte BCDDP (« Breast Cancer Detection and Demonstration Project ») qui regroupe 2 852 cas de cancer du sein et 3 146 témoins indemnes, âgés de 35 à 79 ans, parmi les 284 780 sujets de la cohorte. Cette étude ne porte donc pas sur la prédiction du risque en population générale, mais utilise une population restreinte où le nombre de cas de cancer est équivalent au nombre de témoins non atteints. Le tableau 2.2 regroupe les différents facteurs de risque qui ont été retenus en fonction de leur influence sur le risque de cancer du sein et la discrétisation choisie pour chacun de ces facteurs.

Un modèle de régression logistique non conditionnelle [3] est utilisée. Une différence majeure avec le modèle de Cox, est que ce type de modélisation ne permet pas de

3. On parle de régression conditionnelle quand les malades et les non malades sont appariés en fonction de conditions sur les modalités. Une régression est dite non condionnelle quand il n'y a pas d'appariement effectué.

prendre en compte une durée de suivie différente selon les sujets. La probabilité p de développer un cancer du sein, dans un intervalle de temps défini, est définie par la prise en compte de différentes fonctions :

- une fonction, qui décrit le risque de base de développer un cancer du sein, commune à tous les sujets au temps t,
- une fonction qui décrit le risque relatif de développer un cancer du sein pour une femme qui présente un facteur de risque considéré en comparaison d'un groupe de sujet référence sans facteur de risque connu,
- une fonction qui décrit le taux de mortalité autre que par cancer du sein,
- une fonction qui représente la probabilité de survie aux risques compétitifs.

Tableau 2.2 : Facteurs de risque dans le modèle originel de [Gail 89].

Facteur de risque	Modalités
Âge du sujet (**age**)	<50 ; ≥ 50
Âge à la ménarche* (**menarche**)	<12 ; 12-13 ; ≥ 14
Nombre de biopsies (**biop**)	0 ; 1 ; ≥ 2
Âge à la première naissance (**agenai**)	<20 ; 20-24 ; 25-29 (ou nullipare*) ; ≥ 30
Nombre de parents atteints (**kdeg1**)	0 ; 1 ; ≥ 2

(Âges exprimés en années)

Après estimation des paramètres de la régression, un risque de base est calculé sur la totalité des 284 780 sujets de la cohorte, Gail propose d'estimer le risque en calculant le logarithme du rapport de cote sous la forme suivante :

$$\begin{aligned} -0,749\,48 + 0,094\,01\,(\textbf{menarche}) + 0,529\,26\,(\textbf{biop}) + 0,218\,63\,(\textbf{agenai}) \\ +0,958\,30\,(\textbf{kdeg1}) + 0,010\,81\,(\textbf{age}) - 0,288\,04\,(\textbf{biop}\cdot\textbf{age}) \\ -0,190\,81\,(\textbf{agenai}\cdot\textbf{kdeg1})\,. \end{aligned} \quad (2.12)$$

Pour rendre le calcul plus facilement réalisable, Gail propose de convertir les rapports de cote en risques relatifs à partir des coefficients obtenus pour chacun des attributs. Par exemple, le risque relatif d'une femme ayant eu ses premières règles à 12 ans (catégorie 1) par rapport à une femme ayant ses règles avant 12 ans (catégorie 0) sera égal à :

$$\frac{e^{-0,749\,48+(0,094\,01\times 1)}}{e^{-0,749\,48+(0,094\,01\times 0)}} = 1,099\,. \quad (2.13)$$

L'opération est réitérée pour chacun des attributs et le résultat est montré dans le tableau 2.3. Ainsi pour une femme de 40 ans ayant eu ses premières règles à 12 ans, une maladie bénigne du sein ayant nécessité une biopsie et pas d'antécédent de cancer du sein chez un parent au premier degré (mère ou sœur), on calcul le risque relatif suivant :

$$1,099 \times 1,698 \times 1,548 = 2,89\,. \quad (2.14)$$

Tableau 2.3 : Risques relatifs* associés au modèle de [Gail 89] (âges exprimés en années).

Facteur de risque	Risque relatif
menarche	
<12	1,000
12-13	1,099
≥14	1,207
biop	
age < 50	
0	1,000
1	1,698
≥2	2,882
age ≥ 50	
0	1,000
1	1,273
≥2	1,620

Facteur de risque		Risque relatif
agenai	**kdeg1**	
<20	0	1,000
	1	2,607
	≥2	6,798
20-24	0	1,244
	1	2,681
	≥2	5,775
25-29	0	1,548
	1	2,756
	≥2	4,907
≥30	0	1,927
	1	2,834
	≥2	4,169

Cette femme a alors un risque de cancer du sein 2,89 fois supérieur à une femme du même âge ne présentant pas de facteurs de risque. La capacité des modèles à prédire correctement le risque de développer un cancer du sein est rarement mise en avant dans les articles traitant de la conception de score de risque, probablement parce qu'il est difficile de prédire finement le risque de cancer du sein dans la population générale, notamment à cause du faible nombre annuel de nouvelles personnes atteintes. Cette capacité à prédire correctement le risque de survenue d'un événement binaire (cancer ou pas cancer) peut être traduite, lors de la conception d'un score, par la discrimination et la calibration. La discrimination peut être mesurée par l'aire sous la courbe de la fonction d'efficacité du récepteur, communément appelée courbe ROC (pour « Receiver Operating Characteristic »). Cette mesure de l'aire, comprise entre 0,5 et 1,0 permet de caractériser la capacité d'un classifieur à attribuer un score plus élevé à une personne qui sera atteinte par un cancer qu'à une personne qui ne sera pas atteinte. Sa construction est détaillée en partie 2.3.2.1, page 50. La calibration peut être traduite par le rapport du nombre estimé d'événements sur le nombre observé d'événements, voir partie 2.3.3.2, page 55.

Parmi les limites inhérentes à ce type d'étude et soulevées par les auteurs, on note, d'une part, le choix des facteurs de risque qui reste subjectif et qui pourrait avoir été différent s'il avait été réalisé par d'autres experts et, d'autre part, l'adéquation inconnue entre la base de données utilisée et la population cible. Pour répondre partiellement à ces objections, plusieurs études complémentaires ont été menées : d'une part pour valider le modèle sur la population américaine [Costantino 99] (et en particulier sur des sous-populations émigrées aux États-Unis [Matsuno 11]) et sur

des populations d'autres pays [Decarli 06] et, d'autre part, pour intégrer l'effet de facteurs de risque mis en évidence après la parution du premier article, par exemple la densité mammaire [Tice 05, Tice 08]. De tels modèles atteignent une performance en termes de discrimination comprise entre 0,58 et 0,59. Ces valeurs ont été mesurées a posteriori [Rockhill 01], les auteurs des travaux originaux utilisant une mesure de performance basée sur la calibration.

Dans le cas d'une utilisation par un non spécialiste en statistique, on relèvera la difficulté d'accès à un modèle de régression logistique prenant en compte les risques compétitifs de décès par d'autres maladies. Cette complexité conduit les auteurs à proposer une version simplifiée de l'outil d'estimation du risque par l'utilisation d'un produit de risques relatifs plus aisé à comprendre, mais moins précis. Il n'en reste pas moins que pour les personnes voulant maîtriser les concepts à la base du score de risque, la compréhensibilité du modèle théorique reste limitée.

Modèle de Barlow : Afin d'améliorer les performances des scores de risque pour le cancer du sein, épidémiologistes ont tenté d'intégrer de nouveaux facteurs ou marqueurs de risque. Une étude importante de ce point de vue est de celle de [Barlow 06] qui a permis de produire un double score intégrant des données de densité mammaire issues de centres américains de dépistage du cancer du sein qui ont fourni 2 884 197 radiographies mammaires et des informations, parfois parcellaires, sur les femmes ayant subi ces mammographies*. Après lecture des mammographies par un radiologue qui classe subjectivement la densité des seins en quatre catégories selon la méthode BI-RADS (de 1 : peu dense à 4 : très dense) décrite par [Reston 03], les données de densité ont été intégrées dans un modèle non conditionnel de régression logistique aux côtés des autres facteurs de risque à disposition. Il en résulte la création de deux scores de risque :

- un score pour les femmes non ménopausées qui inclut l'âge, la densité mammaire, les antécédents de cancer du sein et le fait d'avoir subi une biopsie,
- un score pour les femmes ménopausées qui inclut l'âge, la densité mammaire, les antécédents de cancer du sein, le fait d'avoir subi une biopsie, l'indice de masse corporelle, les origines ethniques, le type de ménopause (naturelle ou artificielle), la prise d'un traitement hormonal de la ménopause et la présence d'un éventuel faux positif à la dernière mammographie.

Ses performances en matière d'aire sous la courbe ROC dépendent du modèle considéré. Le modèle pour les femmes non ménopausées affiche une aire sous la courbe ROC de 0,631 tandis que le modèle dédié aux femmes ménopausées a une aire sous la courbe de 0,624.

Pour offrir une meilleure vue d'ensemble des attributs utilisés dans les scores de risque que nous avons évoqué, nous avons regroupé dans le tableau 2.4, page 37, les différents facteurs de risque utilisés selon les différents scores présentés dans cette partie. On y retrouve la nette différence d'approche entre scores familiaux et scores basés sur les facteurs individuels.

Tableau 2.4 : Facteurs pris en compte dans les modèles présentés.

	Claus	Shattuck	Gail	Barlow	BODICEA	IBIS
Facteurs de risque individuels						
Âge	x		x	x	x	x
Âge aux premières règles			x			x
Âge à la ménopause				x		x
Parité*			x			x
Âge à la première naissance			x			x
Indice de masse corporelle				x		x
Biopsie mammaire			x	x		
Antécédent d'hyperplasie* atypique			x			x
Antécédent de carcinome in situ			x			x
Densité mammaire				x		
Traitement hormonal				x		
Faux positif à la mammographie				x		
Facteurs de risque familiaux						
Cancer du sein, âge	x	x			x	x
Nombre d'antécédents familiaux	x	x	x	x	x	x
Au premier degré, âge	x			x	x	x
Au second degré, âge	x				x	x
Au troisième degré, âge					x	x
Âge des apparentés indemnes					x	x
Cancer de l'ovaire, âge		x			x	x
Cancer du sein chez l'homme, âge			x		x	
Bilatéralité du cancer du sein, âge		x			x	x
Cancers multiples		x			x	x
Cancer de la prostate, âge					x	
Cancer du pancréas, âge					x	
Effet cohorte					x	
Présence mutation BRCA 1 ou 2					x	x

2.2.2 Évaluation du risque avec des méthodes de fouille de données

À notre connaissance, il n'existe pas de score de risque du cancer du sein en prévention sélective qui ait été mis au point à l'aide de méthodes de fouille de données. En revanche, plusieurs études ont porté sur le risque de rechute et sur le risque de mortalité après un cancer du sein. Même si les problématiques posées aux auteurs de ces études étaient différentes des nôtres, il est intéressant de les analyser au regard des contraintes qui se posent lors de la création d'un score de prévention en population générale. Parmi ces contraintes, on trouve notamment le type de données utilisées, le déséquilibre dans la distribution des valeurs de l'attribut à prédire dans les données, la compréhensibilité des méthodes d'évaluation utilisées et l'adaptation des scores au contexte d'utilisation.

2.2.2.1 Problématiques abordées et données utilisées

Les études existantes concernant le cancer du sein dans le domaine de la fouille de données ne portent, à notre connaissance, que sur le risque de rechute ou de mortalité après cancer. L'étude d'un tel risque se fait sur des données dans lesquelles l'attribut cible à prédire est équitablement réparti parmi les exemples d'apprentissage.

C'est par exemple le cas des études de [Delen 05] et [Bellaachia 06] qui visent à mesurer la capacité de différents algorithmes de fouille de données à prédire la survie après un cancer du sein. Les données de la SEER (Surveillance Epidemiology and End Results) sont utilisées. Ces données sont collectées et mises à disposition par le NCI. Elles couvrent 28 % de la population américaine.

Pour Bellaachia *et al.*, les données incluent quatre attributs environnementaux (comme l'âge ou le statut marital) et douze attributs décrivant la biologie ou l'avancement de la tumeur (stade de développement ou type histologique), puisque, contrairement aux scores issus de l'épidémiologie, un cancer du sein a déjà eu lieu. Pour Delen *et al.*, après nettoyage et préparation des données, on constate que l'attribut cible est équitablement réparti dans les données puisque 46 % des individus sont de classe positive (*survie*), quand la classe négative (*décès*) est représentée par 54 % des individus.

Les données disponibles sont également recueillies au cours d'études cas-témoin, qui visent à associer à un individu atteint par la maladie, un individu de profil comparable qui n'a pas été atteint par la maladie. Utiliser de telles données conduit également à obtenir un attribut cible réparti de manière équitable parmi les exemples du jeu d'apprentissage. C'est par exemple le cas des données utilisées par [Cong 11] qui permettent de mesurer l'efficacité, en termes de survie, selon le traitement utilisé contre le cancer du sein.

Les problématiques évoquées dans ces études sont donc différentes des problématiques à l'origine des études menées dans le domaine des scores en épidémiologie, puisque c'est essentiellement la survie ou la rechute après un premier cancer qui est

étudiée. De plus, concernant les données utilisées, indépendamment du fait qu'un score de risque en prévention primaire ne donne pas accès à ce type de données puisque le cancer n'a pas encore eu lieu, disposer de mesures biologiques concernant un individu dont on ne sait pas encore s'il est à risque ou pas, est rare. Imposer l'obtention d'une telle mesure biologique rend donc l'utilisation du score moins systématique (voir les coûts d'obtention, partie 3.1.2.2, page 62). Ce constat sur les données est d'ailleurs partagé avec les scores basés sur des méthodes statistiques, par exemple pour le risque cardio-vasculaire, dans lequel des mesures de cholestérol sont utilisées (voir, par exemple le modèle d'Anderson, page 25).

2.2.2.2 Algorithmes utilisés pour mesurer le risque de cancer du sein

Les auteurs étudiant la capacité de prédiction d'algorithmes de fouille de données ont testé de nombreuses méthodes.

On retiendra par exemple, l'étude de [Delen 05] qui permet la comparaison des performances de prédiction : arbre de décision, régression logistique et réseau de neurones sont utilisés pour modéliser la mortalité après cancer du sein. Les performances sont mesurées en matière de sensibilité, de spécificité et de précision car les données ne sont pas déséquilibrées. C'est l'arbre de décision, avec un algorithme de type C5, qui permet la meilleure prédiction avec une précision de 93,6 %.

Une autre étude, menée par [Endo 08], a permis de comparer la performance de sept algorithmes de fouille de données pour prédire le taux de survie à cinq ans après un cancer du sein. Le spectre des algorithmes testé étant large, il est intéressant d'étudier les résultats obtenus (voir tableau 2.5).

Tableau 2.5 : Performances de prédiction de la survie au cancer du sein sur les données SEER d'après [Endo 08].

	Sensibilité	Spécificité	Précision
Régression logistique	97 %	36,3 %	85,8 %
Arbre de décision (C4.5)	97,1 %	34,7 %	85,6 %
Arbre de décision (ID3)	91,6 %	40,9 %	82,3 %
Arbre de décision et Bayésien naïf	92,7 %	46,4 %	84,2 %
Réseau de neurones	92,2 %	50,9 %	84,5 %
Bayésien naïf	92,3 %	47,1 %	83,9 %
Réseau bayésien	91,6 %	40,9 %	82,3 %

Les meilleurs résultats en précision et en sensibilité sont obtenus avec une régression logistique, devant les arbres de décisions de type C4.5 [Quinlan 93] ou ID3 [Quinlan 86] et le réseau de neurones utilisé. Nous n'observons pas de séparation entre les performances des algorithmes explicables visuellement comme les arbres de décision et les méthodes difficilement compréhensibles pour des patients ou des médecins, comme la régression logistique ou le réseau de neurones.

Sur des données similaires, mais préparées différement, [Bellaachia 06] montre des résultats cohérents puisque c'est un arbre de décision C4.5 qui obtient de meilleurs résultats de précision devant le réseau de neurones utilisé.

Différentes sortes d'algorithmes ont été utilisés pour prédire le risque de survie ou de rechute et il est intéressant de noter que les régressions logistiques (utilisés en épidémiologie) et arbres de décisions (facilement compréhensibles si peu profonds) obtiennent régulièrement de bonnes performances dans la littérature.

2.2.2.3 Compréhensibilité des méthodes d'évaluation

Tout comme les méthodes issues de l'épidémiologie présentées dans ce chapitre, les méthodes utilisées pour prédire le risque de rechute ou de mortalité après un cancer du sein sont difficilement compréhensibles par les utilisateurs des scores, qu'ils soient médecins ou patients.

Par exemple, les méthodes de réseau de neurones utilisés par [Delen 05] ou [Bellaachia 06] peuvent difficilement être expliquées au moment de l'utilisation du score, en revanche les arbres de décisions peuvent être envisagés comme une manière simple de présenter le calcul d'évaluation du risque. C'est notamment le cas de l'étude de [Cong 11] qui propose d'utiliser un arbre de décision simple pour déterminer la survie après cancer en fonction du traitement utilisé. Un arbre de décision dont la profondeur maximale est fixée à trois dont la calibration est mesurée grâce à un test du χ^2.

L'efficacité d'un tel algorithme d'évaluation du risque par utilisation des arbres de décision reste à évaluer sur des données plus déséquilibrées, notamment en fonction de la profondeur, et donc de la compréhensibilité, des arbres générés.

2.2.2.4 Prise en compte du contexte d'utilisation

Nous l'avons vu, les scores issus de l'épidémiologie, souvent basés sur des régressions logistiques, sont peu flexibles à cause d'équations de calcul du risque qui sont remodelées à chaque intégration d'un nouvel attribut pour tenter d'améliorer la prédiction.

Si la difficulté est aussi présente dans le domaine de la fouille de données, certains auteurs prennent en compte la nécessité de proposer un score qui soit adapté à certaines contraintes. Ces contraintes ne sont pas fixées en fonction du contexte d'utilisation du score, mais en fonction d'une volonté de maximisation de la performance.

C'est par exemple le cas de la méthode proposée par [Jerez 05]. La modélisation est effectuée sur des données de 1 035 patientes d'un hôpital de Malaga en Espagne et l'objectif est de prédire une rechute après un premier cancer du sein grâce à un réseau de neurones. Si les données sont équilibrées et comprennent une description biologique de la tumeur, en revanche, les auteurs proposent une double phase de

sélection des attributs pour maximiser performances de l'algorithme de prédiction. Lors d'une première phase, des experts du cancer du sein réduisent à 14 le nombre d'attributs considérés grâce à leur connaissance de l'effet de certains attributs sur le risque de rechute. Parmi les attributs de description du profil de la femme, hors la description biologique de la tumeur du premier cancer du sein, on retrouve les attributs sélectionnés par [Gail 89] (voir page 33). Lors d'une seconde phase, les auteurs utilisent un arbre de décision pour limiter encore la taille du jeu. Les attributs non retenus par l'algorithme pour construire l'arbre sont éliminés de la liste des attributs utilisés en entrée du réseau de neurones.

La méthode d'intervention d'un expert pour limiter le jeu de données est intéressante et utile pour augmenter le niveau de performance à moindre coût de calcul. En revanche, il manque une dimension d'adaptation des attributs choisis au contexte d'utilisation, que ce soit en épidémiologie ou en fouille de données, qui peut expliquer la faible utilisation des scores de risque dans le domaine de la prévention primaire en santé.

En conclusion, nous retenons que les études menées dans le domaine de la fouille de données concernant le cancer du sein ne permettent ni de prendre en compte des données déséquilibrées comme on les trouve en population générale lors de la modélisation d'un risque dans un but de prévention primaire ou sélective, ni de choisir un modèle prenant précisément en compte les besoins de différents intervenants en matière d'attributs retenus, de compréhensibilité de la modélisation ou de performances mesurées pour le modèle retenu.

2.2.3 Visualisation : utilisation des scores de risque

Parmi les raisons qui peuvent expliquer la faible utilisation des scores de risque dans le domaine de la santé, et en particulier des scores de risque de cancer du sein, nous émettons l'hypothèse que les outils existants ne sont pas adaptés aux utilisateurs non spécialistes des modèles de risque et notamment que l'appropriation de l'évolution du risque n'est pas facilitée. Cette partie fait le point sur les besoins de lisibilité du modèle utilisé et d'interaction de l'utilisateur avec l'outil informatique, sur les composants d'interface graphique existants pour construire une interface graphique permettant l'utilisation d'un score de risque et les outils existants dans le domaine de la santé.

2.2.3.1 Besoins

L'interface graphique, qui permet d'utiliser le score de risque, doit être une partie de la réponse à la faible utilisation des scores de risque dans le domaine de la santé. Elle doit permettre de mettre en avant la lisibilité de la méthode de modélisation choisie en expliquant schématiquement son fonctionnement pour convaincre l'utilisateur (médecin ou patient) de son bien-fondé. L'objectif est de faire comprendre à l'utilisateur que l'évaluation du risque affichée est le résultat d'une méthode simple et fiable, pas le résultat d'un calcul type boîte noire.

Pour les contextes d'utilisation envisagés, plusieurs aspects doivent être pris en compte. D'abord, le niveau et le type de compétence de l'utilisateur dans le domaine du cancer du sein peut être variable qu'il s'agisse d'un spécialiste du cancer du sein dans de la clinique du risque, d'un médecin d'un centre de dépistage départemental ou d'une patiente. Ensuite, le temps consacré par l'utilisateur au calcul du risque est réduit. La prise en main de l'outil doit donc être immédiate avec un nombre de paramètres à renseigner qui soit limité et un message délivré qui soit juste et clair.

Enfin, l'interface graphique doit favoriser l'appropriation de l'évolution du risque par l'utilisateur. Une solution est de faciliter l'interaction entre l'utilisateur et l'outil de calcul du risque en permettant un calcul du risque instantané du point de vue de l'utilisateur. L'instantanéité du calcul encouragera l'utilisateur de tester rapidement plusieurs profils de risque (le sien ou une version modifiée du sien) tout en visualisant directement la hausse ou la baisse du risque à l'écran.

2.2.3.2 État de l'art

L'étude des différents composants graphiques qui permettent d'entrer des informations dans le système de calcul du score du risque et de visualiser les résultats de l'évaluation est utile pour répondre aux besoins de rapidité et d'appropriation qui doivent être satisfaits pour favoriser l'utilisation de l'outil qui résulte de la phase de déploiement.

Solutions disponibles pour l'entrée des paramètres : Le moyen le plus universel de saisie d'informations est d'utiliser du texte au moyen d'un champ de texte. Il est peu utilisé dans les outils permettant d'utiliser les scores de risque, probablement parce qu'il est de type ouvert ce qui laisse place à de nombreuses erreurs dans le renseignement des valeurs par l'utilisateur. Habituellement, des composants d'interface permettant des réponses fermées de la part des utilisateurs sont utilisés dans les outils de calcul de risque dans le domaine de la santé. Par exemple, l'école de santé publique de Harvard (voir figure 2.1) fait usage des boutons radio ou des cases à cocher, qui permettent de choisir une réponse parmi d'autres, respectivement plusieurs réponses, pour évaluer le risque de diverses maladies, dont les cancers.

Les listes déroulantes constituent une variante de composant d'interface induisant des réponses fermées qui permettent de choisir un item de réponse parmi d'autres. C'est le choix fait par les concepteurs du National Cancer Institute (NCI) pour recueillir les informations dans leur outil de calcul de risque (voir figure 2.2). Ces informations permettent le calcul du risque de cancer du sein en utilisant une évolution du modèle de [Gail 89] par [Costantino 99] et ses améliorations pour des populations spécifiques [Rockhill 01, Matsuno 11].

Les boutons radio, cases à cocher ou listes déroulantes permettent, sous certaines conditions de présentation, de répondre aux besoins d'instantanéité du calcul du niveau de risque permettant une meilleure compréhension du risque et de son évolution. En revanche, ces composants sont inadaptés lorsque le choix de réponse est élevé, or le calcul d'un score de risque de cancer du sein peut nécessiter l'utilisation de plusieurs types d'âges (âge, âge aux premières règles, âge à la ménopause, âge au premier enfant). En fonction de la discrétisation de telles variables continues, l'utilisation de boutons peut être une solution adaptée si elle est couplée à des composants d'interface graphique qui permettent d'améliorer l'interactivité et d'éviter le problème des longues listes de choix.

C'est le cas du curseur de défilement (traduction du terme anglais « slider » souvent utilisé) qui permet en un seul clic, suivi d'un mouvement de souris sans relâcher le bouton, de parcourir successivement plusieurs réponses possibles à une question. Une action qui aurait nécessité, avec les autres composants d'interface évoqués, un clic (appuyer puis relâcher le bouton) et un mouvement de souris pour chacune des valeurs choisies. L'utilisation d'un curseur de défilement permet à l'utilisateur de tester différentes réponses sans relâcher le bouton de la souris ce qui, à notre avis, facilitera l'appropriation de l'évolution du niveau de risque si celle-ci est traduite à l'écran en temps réel. De plus, ce composant graphique est complètement adapté à l'utilisation de bornes tactiles dans un environnement ouvert : il suffit de faire glisser le curseur de défilement avec son doigt.

Une volonté d'amélioration de ces curseurs de défilement pourrait conduire à l'utilisation d'un composant de type ruban qui permet de gagner une gestion facilitée des catégories, une manipulation directe des éléments de réponses sur lesquels l'interface est centrée. Un exemple est proposé par [Eppler 08] dans le domaine de la gestion du risque et reproduit en figure 2.3.

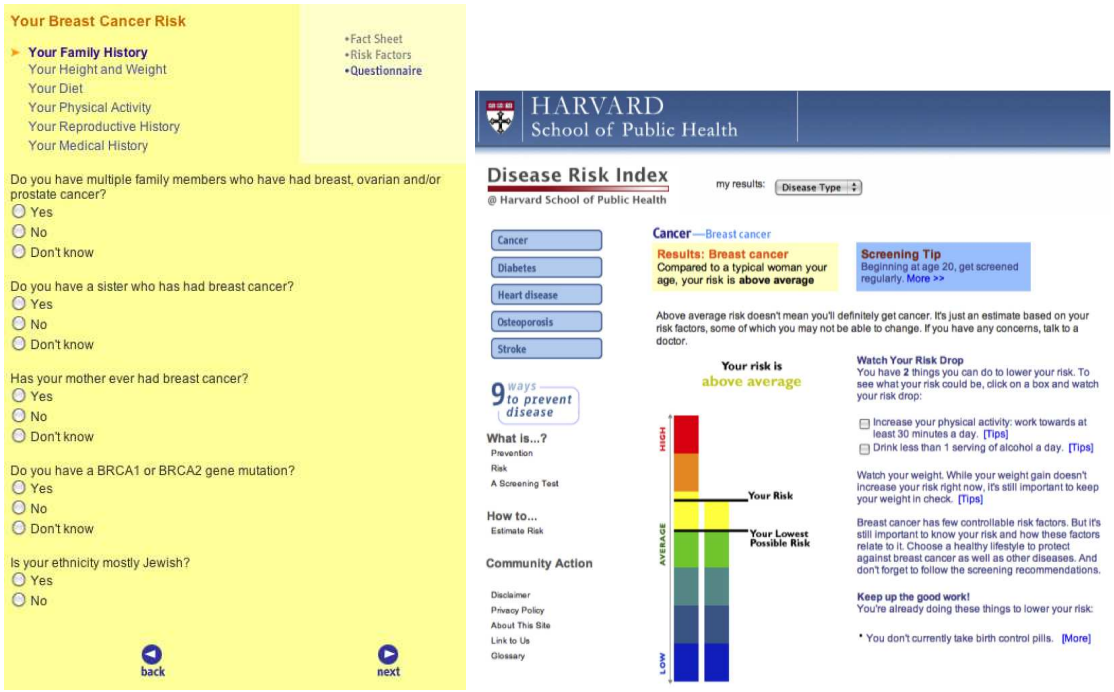

Figure 2.1 : Captures d'écran du site internet de l'École de santé publique d'Harvard présentant un outil de calcul de scores de risque. À gauche, un extrait des questions concernant le cancer du sein. À droite, un extrait d'un résultat d'évaluation de risque pour le cancer du sein [Harvard School 08].

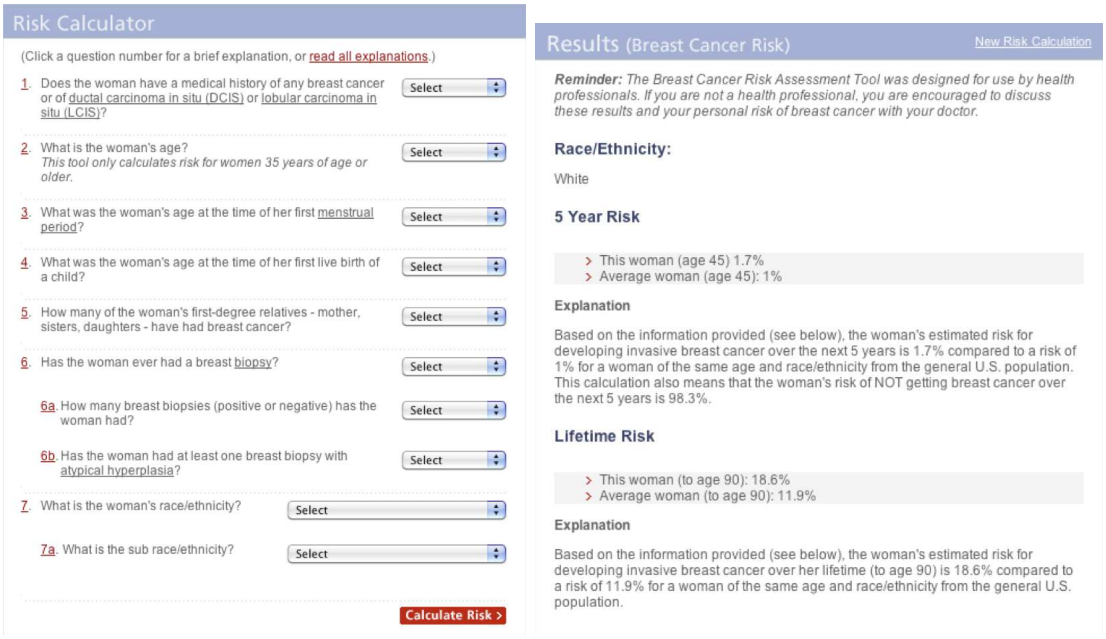

Figure 2.2 : Captures d'écran du site internet du National Cancer Institute présentant un outil de calcul du risque de cancer du sein basé sur le modèle de Gail et ses évolutions. À gauche, un extrait des questions et à droite, un extrait d'un résultat d'évaluation du risque [NCI 11].

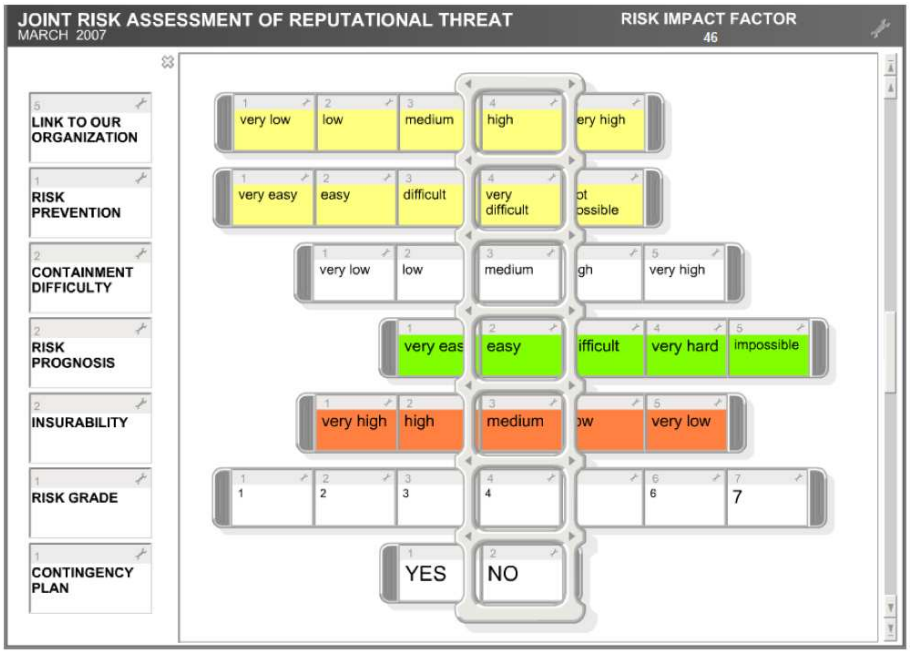

Figure 2.3 : Interface d'entrée d'un outil de gestion du risque basée sur des rubans selon [Eppler 08].

Quelle que soit la solution choisie, elle devra être compatible avec un affichage du niveau de risque qui puisse être mis à jour rapidement en fonction des valeurs renseignées pour favoriser la compréhension de l'évolution du risque affiché. Nous évoquons dans la partie suivante les solutions à disposition pour afficher le niveau de risque calculé.

Solutions disponibles pour l'affichage du niveau de risque : La manière la plus simple d'afficher le niveau de risque, comme pour recueillir les réponses aux questions qui permettent de l'évaluer, est d'utiliser du texte. C'est l'option retenue dans l'outil du NCI fondé sur le modèle de Gail. Un double risque est présenté de manière textuelle : d'une part, un pourcentage d'être victime d'un cancer du sein dans les cinq prochaines années et, d'autre part, un risque sur la durée de sa vie (voir figure 2.2, page 44). Le chiffre est complété par une courte explication du niveau de risque comparé à celui de la population générale et par la probabilité de ne pas être touché par un cancer du sein sur la même période.

Mais le plus souvent les niveaux de risque sont présentés de manière graphique. C'est le cas des résultats de l'outil de l'École de santé publique d'Harvard qui affiche le risque calculé sur une échelle graduée (et colorée en conséquence) aux côtés d'un second risque, le risque minimum atteignable. Peu de facteurs de risque étant modifiables par la femme pour le cancer du sein (voir une analyse de ce problème en partie 3.2.4, page 71), le risque minimum atteignable est donc généralement proche

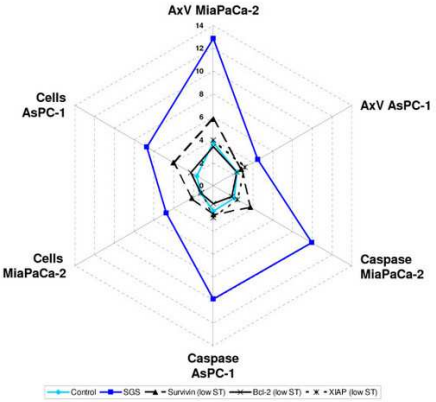

Figure 2.4 : Diagramme de Kiviat (ou diagramme radar) extrait d'une publication en biologie cellulaire [Ruckert 10].

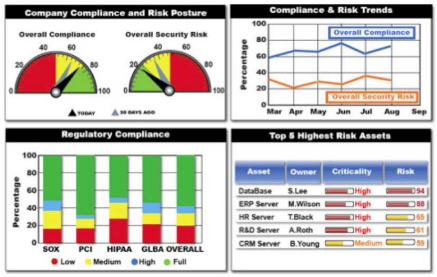

Figure 2.5 : Tableau de bord pour restituer des informations selon [Eppler 08].

2

du risque calculé. Des pistes sont proposées à l'utilisateur pour diminuer le niveau de risque, par exemple (voir figure 2.1, page 44), l'augmentation de l'activité physique et la baisse de la quantité d'alcool ingérée (voir les facteurs de risque du cancer du sein, partie 1.2.3, page 17).

On trouve dans la littérature et dans les différents logiciels, d'autres façons d'afficher un niveau de risque, ou tout simplement un score ou une note en fonction de différents attributs. Prenons deux exemples courants en restitution de résultats dans le domaine de l'interaction homme-machine.

Le diagramme de Kiviat, aussi connu sous le nom de diagramme radar, permet de restituer un niveau de score ou une note (voir la figure 2.4). Un avantage important de ce type de diagramme est que la note ou le niveau affiché est directement fonction de la valeur des attributs représentée sur les branches. Ainsi le niveau affiché est directement proportionnel à la surface générée par les lignes qui relient les différentes valeurs d'attributs sur les branches. En revanche, un tel diagramme peut être difficile à interpréter au vu des nombreuses informations qu'il fournit et peut rebuter l'utilisateur qui n'en n'a jamais rencontré.

De la même manière, le tableau de bord est classiquement utilisé pour afficher de nombreuses informations (voir figure 2.5) à la fois. Mais notre score de risque ne produit pas autant d'information.

Les modèles utilisés pour comprendre l'évolution du risque de cancer et les outils utilisés pour communiquer sur ces modèles cumulent des problèmes tels que l'opacité de la méthode de modélisation ou la difficulté d'interaction de l'utilisateur avec l'outil proposé, notamment en ce qui concerne l'instantanéité de l'outil. Ces défauts sont à la base des propositions que nous effectuons dans le processus présenté au chapitre 3.

2.3 Mesures de la performance

Dans le domaine de la prévention personnalisée pour les grandes maladies, juger de la qualité d'un score de risque nécessite de pouvoir quantifier sa capacité à discerner les personnes à haut risque du reste de la population à risque faible. Dans cette partie, nous expliquons qu'il existe principalement deux moyens de caractériser la performance d'un score de risque : la discrimination et la calibration. Nous détaillons les méthodes que nous utilisons pour mesurer ces deux types de performance. Enfin, nous terminons en expliquant comment nous comparons des classements d'individus réalisés par deux scores différents en complément de la comparaison des mesures de performance propres à chaque score.

2.3.1 Quelle performance mesurer ?

Une manière classique de mesurer la performance d'un classifieur dans le cadre d'un problème de prédiction à deux classes est l'utilisation de la matrice de confusion. Dans le cas d'un score de risque en santé, ces deux classes sont *malade* pour « être atteint par la maladie » d'une part et *sain* pour « ne pas être atteint par la maladie » d'autre part.

La matrice de confusion (également appelée tableau de contingence) est habituellement utilisée en classification automatique lorsqu'un seuil est fixé pour le score au-delà duquel une certaine classe sera attribuée par un modèle et en dessous duquel l'autre classe sera attribuée. Par exemple, si un score est compris dans l'intervalle $[0\,;1]$, on peut fixer un seuil à 0,5 c'est-à-dire en dessous la classe négative est attribuée, au-dessus la classe positive est attribuée. Cette matrice de confusion permet de résumer en un simple tableau, combien d'individus ont été classés comme *malade* à tort (faux positif) et à raison (vrai positif), et, combien d'individus ont été classés comme *sain* à tort (faux négatif), et à raison (vrai négatif), voir figure 2.6.

La matrice de confusion permet de calculer la sensibilité :

$$\text{Sensibilité} = \frac{|\text{Vrais Positifs}|}{|\text{Positifs}|} = \frac{|\text{Vrais Positifs}|}{|\text{Vrais Positifs}| + |\text{Faux Négatifs}|}\,,$$

ainsi que la spécificité :

$$\text{Spécificité} = \frac{|\text{Vrais Négatifs}|}{|\text{Négatifs}|} = \frac{|\text{Vrais Négatifs}|}{|\text{Faux Positifs}| + |\text{Vrais Négatifs}|}\,.$$

L'utilisation d'une matrice de confusion, dont découlent les mesures de sensibilité et de spécificité, implique l'attribution d'une classe à un exemple d'un ensemble de validation. Or un score de risque consiste à calculer la probabilité d'appartenance d'un exemple à une classe donnée (*malade* par exemple) sans attribuer formellement cette classe.

En effet, nous ne pouvons pas attribuer la classe *malade*, et son opposée, à des individus sur un sujet aussi sensible que les maladies car, pour plusieurs raisons, les

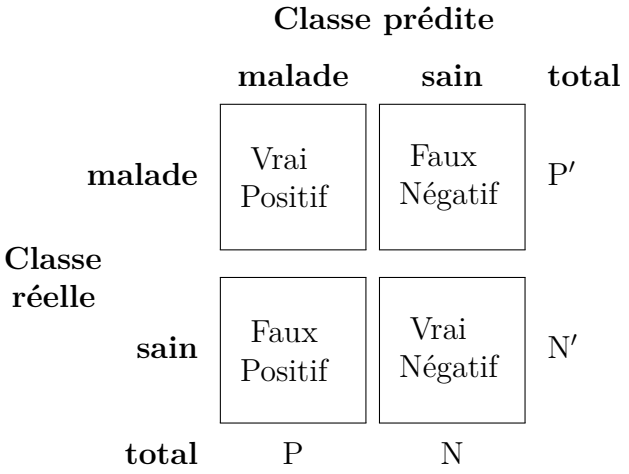

Figure 2.6 : Matrice de confusion utilisée dans le domaine de la santé.

niveaux de réussite de telles prédictions ne sont pas suffisants. Parmi ces raisons, on trouve notamment le peu d'attributs à disposition pour prédire le risque (au vu des coûts d'obtention de différentes natures pour les attributs ou de la connaissance incomplète de la maladie) et le grand déséquilibre de classes dans les données qui conduit les algorithmes à classer tous les individus dans la classe *sain*. Ces raisons sont détaillées dans la partie 3.2.2, page 68.

Un premier moyen de mesurer la performance d'un score de risque, pour lequel aucun seuil n'est fixé pour l'attribution d'une classe, est de mesurer la capacité d'un score à séparer les individus de la classe *malade* des individus de la classe *sain* en attribuant des scores élevés à la première catégorie et des scores significativement moins élevés à la seconde : c'est la discrimination.

Un moyen de compléter cette mesure est de déterminer dans quelle mesure un risque prédit est proche du risque réel [Guessous 10], c'est la calibration.

Le schéma de la figure 2.7, permet d'illustrer les deux types de performance qui seront évalués pour juger de la qualité d'un score de risque. Dans ce schéma, le diamètre du disque correspond au niveau de risque prédit par le score pour un exemple de l'ensemble de validation. La position du disque sur l'échelle [0 ; 1] correspond au niveau de risque réel du même exemple.

- Le cas 1 du schéma présente un score bien discriminant car les profils sont nettement séparés sur l'échelle du risque réel. Il est en revanche mal calibré car le risque évalué n'augmente pas en fonction de la hausse du risque réel.
- Le cas 2 du schéma est au contraire bien calibré car l'augmentation du risque évalué est liée à l'augmentation du risque réel, mais il est mal discriminant car il n'existe pas de nette séparation en fonction du risque réel.
- Enfin, dans le cas 3, le score est à la fois bien calibré et bien discriminant car

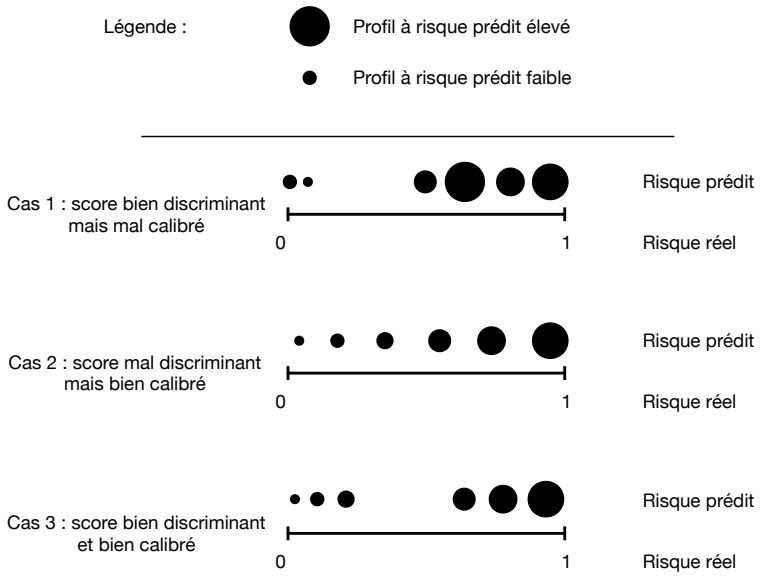

Figure 2.7 : Schématisation de la différence entre calibration et discrimination selon [Guessous 10].

la séparation entre les profils est nette et l'augmentation du risque évalué est liée à celui du risque réel.

Il faut donc utiliser des mesures qui permettent d'évaluer les deux types de performance que sont la discrimination et la calibration.

Pour mesurer la discrimination, nous avons choisi d'utiliser l'aire sous la courbe de la fonction d'efficacité du récepteur plus communément appelée courbe ROC pour « Receiver Operating Characteristic » dont le fonctionnement est détaillé en partie 2.3.2.1, page 50. Cette mesure peut être complétée par l'utilisation du risque relatif observé qui permet de comparer les niveaux de risque chez les individus de haut score et chez les individus de bas score. Ce dernier est plus connu sous son appelation anglaise ORR pour « Observed Relative Risk » et détaillé en partie 2.3.2.2, page 53.

Pour mesurer la calibration, nous avons choisi d'utiliser un diagramme de fiabilité qui permet de représenter l'évolution du risque prédit en fonction du risque réel, voir partie 2.3.3.1, page 54. Ce diagramme sera souvent résumé par l'utilisation du rapport E/O (pour « nombre de cas estimé » / « nombre de cas observé ») qui détaillé en partie 2.3.3.2, page 55, et qui offre une version synthétique du diagramme de fiabilité.

2.3.2 Mesure de la discrimination

2.3.2.1 Aire sous la courbe ROC (AUC)

La courbe de la fonction d'efficacité du récepteur, que nous appellerons courbe ROC dans la suite du manuscrit, a longtemps été utilisée dans le traitement du signal pour détecter les fausses alarmes sur les radars aériens [Egan 75]. Son utilisation a ensuite été étendue à d'autres domaines, dont la conception d'outils de diagnostic dans le domaine médical [Fawcett 06].

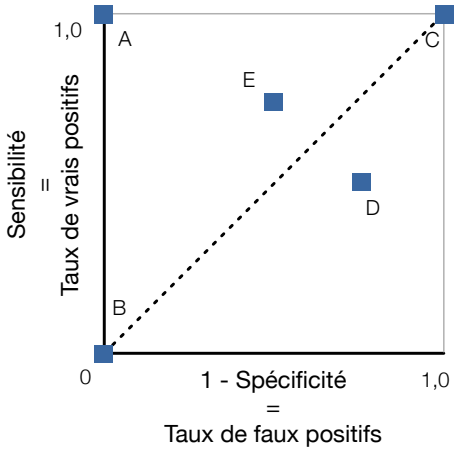

Figure 2.8 : Exemple d'espace ROC.

Espace ROC : La courbe ROC se trace dans un espace en deux dimensions et représente l'évolution de la sensibilité (le taux de vrais positifs) en fonction de 1-spécificité (le taux de faux positifs) comme présenté en figure 2.8. Certains des points de cet espace sont notables :

- le point A(0 ;1) représente la performance parfaite d'un classifieur binaire : pas de faux positifs, uniquement des vrais positifs (les malades prédits comme malades),
- le point B(0 ;0) représente la performance d'un classifieur qui ne classe aucun individu comme malade, tous sont prédits comme sains pour éviter les faux positifs,
- le point C(1 ;1) représente un classifieur qui classe tous les individus comme malades, pour éviter de manquer un malade réel, c'est une forme de minimisation à l'extrême des faux négatifs,
- les points situés sur la diagonale pointillée $y = x$ représentent la performance de classifieurs qui produisent autant de vrais positifs que de faux positifs, leur performance est donc équivalente à attribuer la classe *sain* ou *malade* au hasard,

– les points situés en dessous de cette diagonale ont un taux de faux positifs plus élevé que le taux de vrais positifs, ils sont donc des classifieurs moins bons que le hasard. Puisqu'il n'existe pas de classifieur pire que le hasard, il suffit d'inverser la classe prédite pour replacer la performance au-dessus de la diagonale de l'espace ROC. C'est le cas du point D qui représente la performance d'un classifieur dont les prédictions sont l'exact inverse du classifieur dont la performance est représentée par le point E [Flach 03].

Le choix d'un classifieur peut donc se faire selon que l'on souhaite minimiser le taux de faux positifs ou de faux négatifs.

Courbe ROC : Pour construire un score de risque, il n'est pas nécessaire de choisir un seuil qui permette de déterminer la classe à attribuer, il n'est donc pas possible de construire de matrice de confusion pour évaluer la sensibilité et la spécificité du score. Le principe de la courbe ROC est d'utiliser le score pour attribuer un risque à différents individus et de ne pas choisir un seuil de prédiction, mais autant de seuils possibles de plus l'infini à moins l'infini, pour produire autant de matrices de confusion que de seuils choisis et donc autant de points dans l'espace ROC.

Tableau 2.6 : Exemple de scores attribués à 20 individus malades (m) ou sains (s) (inspiré de [Fawcett 06]).

Indiv.	Classe	Score	Indiv.	Classe	Score
1	m	0,90	11	m	0,38
2	m	0,80	12	s	0,37
3	s	0,74	13	m	0,30
4	m	0,70	14	s	0,28
5	m	0,55	15	s	0,25
6	m	0,54	16	s	0,20
7	s	0,50	17	m	0,17
8	s	0,49	18	m	0,13
9	m	0,40	19	s	0,10
10	s	0,38	20	s	0,05

Afin d'illustrer la construction d'une courbe ROC, détaillons un exemple. Si l'on souhaite construire la courbe ROC qui permet d'évaluer la discrimination du score qui a servi pour évaluer le risque des individus du tableau 2.6, il faut générer les matrices de confusion pour des seuils de valeur supérieure à 0,90 jusqu'à une valeur inférieure à 0,05. Ainsi, pour une valeur de seuil supérieure à 0,90, tous les individus sont prédits comme étant de la classe *sain*, les taux de vrais positifs et de faux positifs valent 0, le premier point de la courbe est donc placé à l'origine du repère. Puis, on fait diminuer le seuil à une valeur comprise entre 0,80 et 0,90, ce qui permet de construire la matrice de confusion du tableau 2.7. Le taux de vrais positifs vaut

1/10, soit 0,1, tandis que le taux de faux positifs vaut 0/10, soit 0. Le second point de la courbe ROC est placé en (0 ;1). La suite de la courbe ROC est construite en continuant de faire diminuer le seuil et présentée en figure 2.9.

Tableau 2.7 : Matrice de confusion pour un seuil compris entre 0,80 et 0,90.

	malade prédit	sain prédit	total
malade réel	1	9	10
sain réel	0	10	10
total	1	19	20

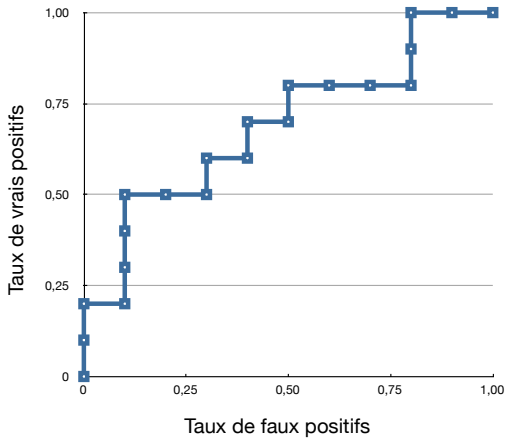

Figure 2.9 : Courbe ROC correspondant aux données du tableau 2.6, page 51.

À la lumière de cet exemple, il est important de noter deux caractéristiques de la courbe ROC. D'abord, cette mesure est indépendante de la calibration : c'est en effet l'ordre des individus classés par niveau de score qui influe sur la position des points de la courbe et pas la valeur du score. Si dans l'exemple présenté en tableau 2.6, page 51, les valeurs peuvent être assimilées à des probabilités comprises dans l'intervalle [0 ; 1], ces valeurs de scores auraient pu être très différentes, du moment que l'ordre des exemples classés par scores est conservé, la courbe reste identique.

Il est également important de noter que la distribution des classes (*sain* et *malade* dans notre cas) n'influe pas sur la courbe ROC. En effet, le calcul du taux de vrais positifs ne fait intervenir que des éléments de la classe des positifs (la classe *malade* réelle de la figure 2.6, page 48) de même que le calcul du taux de faux positifs ne fait intervenir que les éléments de la classe des négatifs (la classe *sain*). Au contraire d'autres mesures de performance qui utilisent des éléments des deux classes réelles et qui mettent donc en jeu le rapport entre l'importance des deux

classes dans le calcul de l'indicateur, la courbe ROC n'est obtenue que par le calcul de rapports à l'intérieur même des classes réelles. Cette mesure de performance est donc indépendante de la distribution des classes au sein des données utilisées.

Aire sous la courbe ROC : Puisque la courbe ROC est un indicateur en deux dimensions, la comparaison de la discrimination des classifieurs nécessite la production d'un indicateur simple à une dimension : c'est l'aire sous la courbe qui est utilisée pour mesurer cette performance [Hanley 82], elle est appelée AUC en anglais pour « Area Under the ROC Curve ». Les valeurs maximales des taux utilisés pour tracer la courbe étant égales à 1,0, la valeur maximale de l'aire sous la courbe ROC est de 1,0. Et, étant donné que les points situés sur la diagonale correspondent à un classifieur attribuant une classe au hasard, la valeur minimale de l'aire sous la courbe est de 0,5.

L'aire sous la courbe a une caractéristique statistique importante puisqu'elle représente la probabilité qu'un classifieur place un exemple positif devant un exemple négatif.

2.3.2.2 Risque relatif observé

Afin de compléter l'aire sous la courbe ROC pour mesurer la discrimination, on utilise le risque relatif observé, abrégé en ORR pour « Observed Relative Risk ». En effet, différentes attributions de classe à des exemples d'un ensemble de validation peuvent conduire à une même courbe ROC et différentes courbes ROC peuvent conduire à une même aire sous la courbe. L'ORR permet d'avoir une vision plus précise de la distribution des classes dans la liste des exemples classés par score.

Lorsque les individus sont classés en fonction du score qui leur est attribué par le classifieur, l'ORR correspond au rapport du nombre de cas observés dans le décile supérieur sur le nombre de cas observés dans le décile inférieur. Ainsi, plus le classifieur place d'exemples de la classe réelle *malade* dans les 10 % d'individus qui ont le score le plus haut et moins il en place dans les 10 % d'individus qui ont les niveaux de risque les plus faibles, alors plus haute (et donc meilleure) est la valeur de l'ORR. Si autant de cas de la maladie sont présents dans les déciles inférieur et supérieur, la valeur de l'ORR sera de 1, la discrimination est mauvaise. Si le décile de score inférieur ne comporte aucun individu de classe *malade* réel, la valeur d'ORR est infinie.

Si la mesure de la discrimination, par l'aire sous la courbe ROC (AUC) et l'ORR, permet de détecter si les individus de la classe *malade* obtiennent en moyenne un score plus élevé que les individus de la classe *sain*, en revanche, elle ne permet pas de dire si le risque prédit est proche du risque réel. L'utilisation de la calibration permet de mesurer ce type de performance.

2.3.3 Mesure de la calibration

La calibration d'un score est la mesure de sa capacité à prédire un niveau de risque proche du niveau de risque réel. Pour la mesurer, nous avons choisi d'utiliser le diagramme de fiabilité qui peut être résumé par le rapport du nombre estimé de cas de maladie sur nombre observé de cas, deux méthodes que nous décrivons dans cette partie.

2.3.3.1 Diagramme de fiabilité

L'objectif du diagramme de fiabilité est de déterminer si le nombre estimé d'individus de la classe *malade* sur une population de test, est proche du nombre de cas de maladie effectivement observé dans cette population. Puisque la création d'un score de risque n'engendre pas d'attribution de la classe *malade*, il n'est pas possible de compter à combien d'individus la classe *malade* a été attribuée.

Le nombre estimé de cas de maladie est donc obtenu en effectuant la moyenne des scores attribués à condition que ces scores expriment une probabilité. Ce nombre estimé de cas de maladie est ensuite comparé au nombre de personnes effectivement malades dans l'échantillon de population utilisé. Cette comparaison est réalisée par quantile de population pour être plus précise.

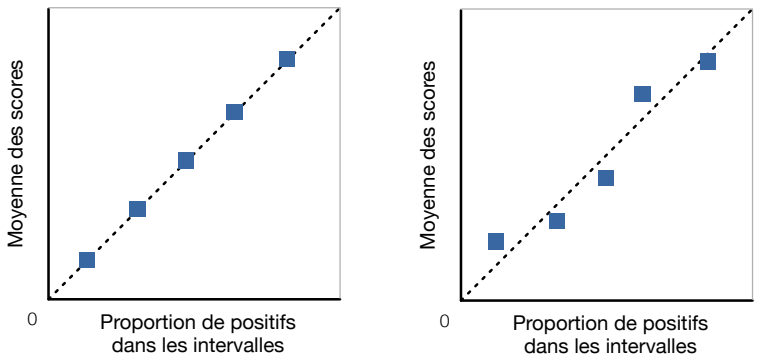

Figure 2.10 : Exemple de diagramme de fiabilité, parfait à gauche, moins bon à droite.

Le diagramme de fiabilité est un graphique dont chaque point représente le niveau moyen du score sur l'axe des ordonnées en fonction de la proportion de classe positive (classe *malade*) pour un intervalle de score, voir 2.10. Lorsque la moyenne des scores est équivalente au nombre d'exemples de classe positive dans chaque quantile, le diagramme prend la forme d'une droite diagonale de type $y = x$, voir côté gauche sur la figure 2.10. En revanche, plus les points s'éloignent de la diagonale, plus la différence entre le nombre d'exemples positifs et la moyenne du score est marquée, voir côté droit sur la figure 2.10.

2.3.3.2 Rapport du nombre estimé de cas de maladie sur le nombre observé de cas de maladie (E/O)

Afin de faciliter la comparaison entre les différents scores produits, le rapport entre le nombre d'exemples de classe positive estimé et le nombre observé d'exemples de classe positive est souvent utilisé. On l'appelle le rapport estimé sur observé, ou E/O en abrégé. Ce rapport peut être calculé sur l'ensemble du jeu de validation ou, pour obtenir une vision plus précise de la calibration, par quantile de score.

Si le score n'est pas exprimé sous la forme d'une probabilité, il doit y être converti pour estimer un nombre d'exemples de classe positive attendu dans chaque quantile. La somme attendue des exemples de positives est ensuite comparée au nombre total d'exemples de classe positive effectivement présente par le calcul du rapport estimé sur observé.

Plus le rapport est proche de 1 (inférieur ou supérieur), plus le nombre d'exemples positives estimé est proche du nombre observé, meilleure est la calibration.

2.3.4 Comparaison avec d'autres scores

La comparaison des niveaux de risque, fournis par un score, comparativement à un autre, est un élément important pour caractériser la performance d'un score. Il est notamment utile de connaître de quelle manière les individus sont classés par deux scores de risque différents.

2.3.4.1 Mesure de corrélation

Une manière de comparer le classement d'individus évalués par deux scores de risque différents est d'utiliser une méthode classique en statistique : le calcul de la corrélation. Elle a pour avantage de représenter un indicateur synthétique.

Nous avons choisi d'utiliser la corrélation de Pearson, symbolisée par ρ (« rho »), qui se définit comme le rapport de la covariance entre deux attributs X et Y (décrits par les valeurs assignées par chacun des scores) sur le produit de leur écart-type σ [Schwartz 93] :

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}\,. \tag{2.15}$$

La valeur du coefficient de corrélation ρ est comprise dans l'intervalle [-1 ; 1] :

- si ρ vaut 0, alors il n'existe aucune corrélation entre les valeurs des attributs et les scores sont tout à fait dissemblables,
- si ρ est positif, les valeurs élevées de X sont associées à des valeurs élevées de Y et, en cas limite, lorsque ρ vaut 1, tous les points (X, Y) sont exactement alignés sur une droite de pente positive,

– si ρ est négatif, le valeurs élevées de X sont associées à des valeurs basses de Y et, en cas limite, lorsque ρ vaut -1, tous les points (X, Y) sont alignés sur une droite de pente négative.

2.3.4.2 Mesure de concordance

Une autre manière de comparer deux scores est de comparer les rangs assignés aux individus par chacun des scores en présence. C'est ce que permet la mesure de concordance qui peut être résumée par un indicateur numérique : le coefficient de concordance nommé κ (« kappa »). Celui-ci nécessite la construction d'une matrice de concordance pour être calculé.

Tableau 2.8 : Matrice de concordance pour n individus classés par tertiles de valeurs de deux scores.

		Score 2			
		tertile 1	tertile 2	tertile 3	Total
Score1	tertile 1	n_{11}	n_{12}	n_{13}	i_1
	tertile 2	n_{21}	n_{22}	n_{23}	i_2
	tertile 3	n_{31}	n_{32}	n_{33}	i_3
	Total	c_1	c_2	c_3	n

Le tableau 2.8 présente la répartition de n individus d'un ensemble de test en fonction des valeurs obtenues par l'utilisation de deux scores de risque différents, les valeurs étant réparties en trois tertiles 1, 2 et 3 pour l'exemple. La suite étant valable pour tout nombre de quantiles, tant que celui-ci reste inférieur ou égal au nombre d'individus. L'effectif d'une case est représenté par le nombre n_{ij}

Dans ce tableau, un accord parfait entre les deux scores pour l'attribution des tertiles est observé s'il n'y a aucun individu en dehors des trois cases (n_{11}, n_{22} et n_{33}) de la diagonale, appelées paires concordantes. Si un effectif d'une autre case est supérieur à 0, alors les individus n'auront pas été classés dans le même tertile par les deux scores. Plus on s'éloigne de la diagonale (n_{13} et n_{31} par exemple), plus la reclassificiation est faite dans un tertile éloigné, et donc plus le classement effectué par les scores a été différent.

Le coefficient κ nécessite le calcul de la concordance observée et de la concordance attendue. La concordance observée est le taux de paires concordantes observées C_O. La concordance attendue C_A est la somme des effectifs théoriques divisée par l'effectif total, soit dans notre cas :

$$C_O = \frac{1}{n}\left(n_{11} + n_{22} + n_{33}\right) \quad \text{et} \quad C_A = \frac{1}{n}\left(\frac{c_1\, i_1}{n} + \frac{c_2\, i_2}{n} + \frac{c_3\, i_3}{n}\right). \tag{2.16}$$

Le coefficient κ est alors défini par :

$$\kappa = \frac{C_O - C_A}{1 - C_A}, \tag{2.17}$$

qui représente le degré d'accord dans les classements obtenus par l'attribution des valeurs par les scores 1 et 2. Ce degré d'accord est compris dans l'intervalle [-1 ; 1] :

- la valeur 0 correspond à un accord dû au hasard,
- la valeur 1 est obtenue en cas d'accord absolu entre les scores (seuls les effectifs des diagonales sont pourvus, donc C_O vaut 1 donc numérateur et dénominateur sont égaux),
- la valeur -1 est obtenue en cas de désaccord absolu entre les deux scores (aucun effectif dans les cases de la diagonale).

Le calcul du κ peut être pondéré pour prendre en compte la distance de reclassification, ainsi un individu reclassé dans le tertile jouxtant le tertile d'origine pèsera moins dans le calcul de l'indice qu'un individu reclassé dans le tertile le plus éloigné. L'intérêt de cette pondération est plus grand quand les individus sont divisés en déciles ou plus. Plus la reclassification est faite dans une classe éloignée, plus l'indice κ est diminué.

Tableau 2.9 : Échelle de Landis et Koch habituellement utilisée pour caractériser l'accord estimé par κ.

Valeur de κ	Interprétation
< 0	Désaccord
[0 – 0,2 [	Mauvais
[0,2 – 0,4 [	Médiocre
[0,4 – 0,6 [	Moyen
[0,6 – 0,8 [	Bon
$> 0{,}8$	Très bon

L'échelle de Landis et Koch, présentée dans le tableau 2.9, est usuellement retenue pour qualifier la concordance [Landis 77].

Dans ce chapitre, nous avons expliqué que les scores de risque dans le domaine de la santé ont tous été construits avec des modèles statistiques et que les premiers d'entre eux ont été basés sur les données de la cohorte américaine de Framingham. Nous avons détaillé les scores de risque existants pour le cancer du sein, maladie à laquelle nous appliquons nos travaux, en pointant leurs faiblesses. Nous avons également analysé les différents outils conçus pour communiquer sur les risques ainsi modélisés. Enfin, nous avons expliqué quelles mesures de performance, exprimée sous la forme de la discrimination et de la calibration, nous utilisons dans ce manuscrit pour qualifier la performance des scores de risque produits.

3 Proposition d'un processus pour la construction d'un score de risque, application au cancer du sein

Améliorer la prévention des maladies par l'utilisation de scores de risque suppose d'être en capacité de produire des scores qui soient à la fois performants dans la détection des profils à risque et adaptés au contexte d'utilisation qui peut fortement varier en fonction des maladies concernées et des spécialités médicales concernées. Par exemple, l'utilisation d'un score de risque de cancer du sein au moment d'une mammographie chez un radiologue permettra l'utilisation d'un ou plusieurs indices de densité mammaire tandis qu'un tel score utilisé chez un médecin généraliste devra utiliser des valeurs qui peuvent être obtenues par un simple questionnement de la patiente.

Afin de simplifier la production de scores de risque qui soient à la fois performants, adaptés à la maladie et adaptés au contexte, nous proposons dans ce chapitre un processus général de construction de score de risque en santé. La description de ce processus est illustrée par le cas de la création d'un score pour le cancer du sein dans deux contextes différents, d'une part un projet d'espace pédagogique sur le cancer du sein qui soit ouvert au grand public et, d'autre part, une clinique du risque qui reçoit des femmes à haut risque de cancer du sein. De ces contextes, nous déduisons des objectifs généraux permettant la conception de scores facilement adaptables au contexte. Nous complétons ces objectifs par l'analyse des problématiques spécifiques aux données médicales.

En tenant compte de ces contextes, objectifs et problématiques, la réalisation d'une proposition de processus de création de scores de risque implique l'intégration de connaissances fournies par les experts, par exemple sur la spécialité médicale ou le contexte d'utilisation envisagé. Sur la base de cette proposition de processus, nous montrons un exemple de création de score de risque sur le cancer du sein. Cet exemple est présenté de la phase de compréhension du contexte métier et des données à la phase de déploiement de l'outil en passant par les phases de modélisation et d'évaluation.

3.1 Objectifs de santé publique et contraintes induites

Pour proposer un processus de création de scores de risque qui soit appliqué au risque de cancer du sein, nous abordons dans cette partie le contexte métier pour expliquer les contextes d'utilisation envisageables pour les scores de risque à

produire. D'une généralisation des contextes décrits, nous déduisons nos objectifs, à la fois en ce qui concerne le processus de création de scores de risque et l'algorithme à utiliser pour faciliter la compréhension du calcul du risque.

3.1.1 Analyse des contextes d'utilisation envisagés

Parmi les différents contextes d'utilisation envisageables pour un score de risque du cancer du sein ayant pour but d'améliorer la prévention en population générale, nous décrivons deux types de contexte d'utilisation de scores de risque : d'une part un projet de clinique du risque à long terme et, d'autre part, un projet plus simple d'espace pédagogique sur le cancer.

3.1.1.1 Projet clinique du risque

Le principe général du projet de clinique du risque est présenté en partie 1.1.2, page 9. L'objectif de ce projet est de suivre les femmes les plus à risque dans un centre de lutte contre le cancer au cours d'une consultation d'une journée pendant laquelle les examens nécessaires sont réalisés, et les résultats analysés, en un même lieu. L'objectif est d'accélérer le processus qui peut habituellement s'étaler sur plusieurs semaines entre le dépistage initial et l'éventuelle mise sous traitement après une biopsie ayant permis la détection d'un cancer du sein.

Pour permettre la mise en place d'une telle prise en charge en fonction du risque, il faut disposer de scores de risque qui permettent d'évaluer de manière efficace le niveau de risque d'une femme. Une possibilité est d'associer les centres de dépistage départementaux au projet de clinique de risque : les femmes pourront y obtenir une évaluation de leur risque au moment de la mammographie de contrôle et les femmes les plus à risque se verront proposer un suivi plus poussé en fonction de leur niveau de risque ou d'une suspicion de cancer à l'analyse du cliché mammographique.

Une seconde possibilité d'utilisation d'un score de risque dans un tel projet de clinique du risque est l'information des patientes le jour de la consultation en une journée dans un centre de lutte contre le cancer. Les patientes chez qui un cancer du sein n'est pas détecté, suite à une mammographie décelée comme suspecte, sont conseillées sur la manière de se comporter vis-à-vis du risque de cancer du sein et se voient proposer des solutions de dépistage. Un outil de mesure du niveau de risque peut donc être utile afin d'adapter les conseils prodigués.

3.1.1.2 Projet d'espace pédagogique Hygée

L'utilisation des scores de risque est également envisagée afin de mener des actions de prévention et de faciliter la communication avec les femmes. À ce titre, l'exemple du projet d'espace pédagogique-prévention du centre Hygée est intéressant.

En effet, dans le cadre de l'axe *Santé publique* du Plan Cancer en Rhône-Alpes et en Auvergne, le pôle hospitalo-universitaire de Saint-Étienne a été retenu pour construire une plate-forme appelée « Centre régional de ressources pour l'information, la prévention et l'éducation sur les cancers », ou « Centre Hygée ». Un des projets de ce centre est de construire un espace d'exposition pédagogique ouvert au public et permettant la mise en scène d'une information scientifique validée correspondant aux demandes du public sur l'état des savoirs en cancérologie, pour ce qui est des données épidémiologiques, des facteurs de risque, de la prévention, de la détection précoce ou de l'actualité des traitements et de la recherche.

Un outil informatique permettant de calculer de manière instantanée un niveau de risque de cancer du sein pourrait ainsi être mis à disposition du public, par exemple sous la forme d'une borne informatique, pour répondre à l'objectif affiché de mettre l'accent sur le visuel et l'interactivité. Les utilisateurs pourraient ainsi découvrir l'impact des facteurs de risque sur le cancer du sein. L'interactivité permettrait l'appropriation de la manière dont le risque évolue en fonction des facteurs de risque renseignés et de l'évolution de ceux-ci s'ils peuvent être modifiés. Par exemple, l'utilisateur pourrait visualiser un risque à la baisse avec un traitement hormonal substitutif différent ou un niveau d'activité physique plus élevé.

Si une telle utilisation de bornes interactives est envisageable en clinique du risque, ce projet a pour différence majeure de ne pas intégrer de personnel médical dans l'information de prévention qui est délivrée aux visiteurs. En revanche, les objectifs d'information sur les facteurs de risque de cancer du sein et leurs impacts sur le risque, montrés à travers des outils efficaces utilisant des méthodes compréhensibles par tous, sont identiques.

La description de ces deux contextes met en avant la diversité des publics qui seront amenés à utiliser un score de risque : du patient qui n'a pas de connaissance précise des facteurs de risque du cancer du sein, au médecin, spécialiste de la maladie qui souhaite évaluer le risque d'une femme pour adapter la prévention proposée. Ces exemples montrent également que le coût d'acquisition des valeurs des attributs varie en fonction du contexte d'utilisation. L'impact de ces contraintes est discuté dans la partie suivante ou nous détaillons les contraintes induites par ces objectifs.

3.1.2 Prise en compte des contraintes liées au contexte

Notre étude, d'une part, des points forts et des points faibles des scores de risque existants et des processus de modélisation (voir chapitre 2) et, d'autre part, des contextes métiers envisagés pour l'utilisation des scores produits, nous amène à énoncer les contraintes qui pèsent sur nos objectifs. Elles portent sur le processus de conception des scores de risque, sur les modèles de prédiction utilisés et les résultats de ces scores.

3.1.2.1 Contraintes sur le processus permettant la prise en compte du contexte par l'interaction avec l'expert

Une part des critiques qui sont faites aux scores de risque (principalement dans les articles de revue ou les rapports des autorités de santé) se concentre sur l'inadéquation du score proposé au contexte d'utilisation envisagé [ANAÉS 04]. Il faut donc proposer un processus simple et flexible qui permette de prendre en compte le contexte d'utilisation au plus tôt dans la construction du score. Le processus doit pouvoir être reparcouru facilement pour pouvoir adapter aisément un score à un autre contexte d'utilisation, par exemple en fonction de spécialités médicales différentes pour une même maladie, ce qui implique notamment la possession d'attributs (décrivant des facteurs de risque) différents.

Le processus de construction du score de risque doit permettre de faciliter le choix des attributs qui seront retenus pour prendre en compte les facteurs de risque de la maladie. Tous les facteurs ne sont pas utiles, ni même disponibles, dans tous les contextes d'utilisation du score de risque, que ce soit, par exemple, en médecine de ville, en consultation dans un service spécialisé d'hôpital ou lors d'un examen de dépistage de routine. Le processus doit permettre d'adapter les attributs utilisés au contexte d'utilisation du score de risque notamment pour améliorer l'acceptabilité du score par les utilisateurs (qu'ils soient médecins, spécialiste de la maladie ou non, ou patients). Ceux-ci peuvent en effet avoir des a priori sur les attributs adéquats à utiliser dans un score de risque et une contrainte est de permettre que le processus permette de prendre en compte ces a priori.

Une autre contrainte est de permettre que le processus de création du score de risque autorise l'utilisation de données pour lesquelles le taux de valeurs manquantes est variable. De plus, le processus doit permettre de gérer le cas d'attributs dépendants les uns des autres. Par exemple, la prise d'un traitement peut être conditionnée à la réalisation d'un événement (alerte cardiaque, ménopause, etc.). Il faut donc permettre que certaines valeurs restent manquantes dans les données utilisées quand l'événement conditionnel n'est pas intervenu.

Lors du processus de conception du score de risque, il faut également prendre en compte le contexte informatique d'utilisation de ce score. Le temps de calcul nécessaire à l'utilisation du score et donc la puissance de calcul disponible ou la possibilité d'accéder à un serveur distant et donc la présence d'une connexion internet, sont des paramètres à considérer absolument lors du processus de conception du score de risque, car ils peuvent avoir une influence sur le type de modélisation des données et le mode de calcul du score qui seront retenus.

3.1.2.2 Contraintes sur le modèle d'évaluation et compromis sur les performances

Si des contraintes ont été identifiées pour le processus de création de scores de risque, nous souhaitons également influer au niveau du modèle choisi pour calculer un niveau de risque. Nous avons donc également des contraintes à ce niveau.

La première contrainte est d'atteindre un haut niveau de performance du score pour sa capacité à détecter les profils les plus à risque dans une population donnée. Il ne s'agit pas forcément de maximiser l'efficacité du score de risque, mais de permettre le choix d'une combinaison d'attributs qui intègre le facteur performance pour réaliser un compromis en regard d'autres contraintes qui peuvent dégrader la performance dans des proportions plus ou moins importantes.

Parmi ces contraintes qui peuvent nécessiter d'effectuer un compromis entre adéquation et performance, on trouve notamment la compréhensibilité du modèle utilisé pour détecter les profils des personnes les plus à risque. En effet, nous pensons que la faible utilisation des scores de risque dans le domaine médical, et pour le cancer du sein en particulier, est en partie due à l'incompréhension par les utilisateurs des modèles théoriques utilisés pour prédire le risque.

D'autres éléments peuvent également impacter l'objectif de performance dans la détection des profils à risque : le coût d'acquisition des éléments qui servent à estimer le risque doit être pris en compte puisque nous sommes dans un contexte de prévention sélective. Cela implique que les populations qui peuvent bénéficier de telles méthodes de mesure du risque peuvent être très larges, entraînant des coûts élevés à mettre en rapport avec les bénéfices retirés. Nous avons identifié quatre types de coût d'acquisition qu'il est nécessaire d'évaluer avant de choisir d'intégrer un attribut dans la combinaison de facteur retenue pour le score.

- Coût financier : l'obtention de certains facteurs de risque peut nécessiter de recourir à des examens coûteux financièrement comparativement au gain de performance obtenu par leur intégration dans le score de risque.
- Coût en risque : l'obtention de certains facteurs de risque peut nécessiter de recourir à des examens qui peuvent impliquer un risque, même minime sur la personne qui le subit, par exemple une radiographie.
- Coût en temps : certains facteurs de risque ne sont pas décelables à l'œil nu et peuvent nécessiter des examens approfondis qui font que le risque ne pourra pas être estimé au moment nécessaire. Par exemple un examen sanguin peut ne pas spécialement coûter cher, ni être risqué, mais décaler le calcul du score dans le temps, voire l'empêcher.
- Coût émotionnel : l'obtention de certains facteurs de risque peut impliquer des examens peu à très désagréables pour le patient. Par exemple le niveau d'invasivité d'un examen.

Pour toutes ces raisons, il est nécessaire que le score de risque inclut des facteurs de risque dont l'utilisation a été évaluée en fonction du contexte d'utilisation et de l'objectif d'utilisation.

En supplément des contraintes décrites pour le processus de création de scores de risque et de la méthode de modélisation du risque, la construction d'un score de risque dans le domaine de la santé en population générale implique de prendre en compte diverses problématiques liées à ces données.

3.2 Problématiques liées aux données de santé

En utilisant l'exemple du cancer du sein nous expliquons dans cette partie le mode de recueil des données épidémiologiques en population générale et les problèmes qui peuvent découler de l'utilisation de telles données pour la construction d'un score de risque en santé : déséquilibre du rapport entre le nombre de personnes saines et le nombre de personnes malades qui complexifie la tâche de prédiction, prise en compte de la confidentialité des données dans un contexte de santé et prise en compte des possibilités d'actions affichées en fonction des attributs choisis pour calculer le score.

3.2.1 Recueil des données utilisables pour construire un score de risque

Avant d'aborder les problématiques liées aux données de santé, une présentation globale de l'origine des données de cancer du sein que nous utilisons est nécessaire. Cette présentation permet de mettre en perspective les forces et les faiblesses des données à disposition, car une modélisation pertinente dépend notamment de la qualité des données à disposition, par exemple en matière de fraicheur, d'exhaustivité ou d'exactitude des données [Akoka 07]. Nous décrivons d'abord l'origine et le recueil des données d'une base américaine disponible publiquement afin de confronter nos résultats à ceux de la communauté scientifique, à la fois dans le domaine de l'épidémiologie dont sont issus la majorité des scores de risque en santé et dans le domaine de la fouille de données dont est issu l'algorithme des plus proches voisins que nous expérimentons dans le chapitre 5. Puis nous décrivons le recueil des données d'une étude de cohorte française afin de permettre la conception d'un score de risque adapté aux femmes françaises.

3.2.1.1 BCSC, une base américaine publique, mais limitée

Aux États-Unis, le « Breast Cancer Surveillance Consortium » (BCSC) est issu de la volonté des autorités publiques d'une part de mesurer la qualité et l'impact des programmes de dépistage du cancer du sein et, d'autre part, de constituer une source de données à des fins de recherche sur les causes et les conséquences du cancer du sein [Ballard-Barbash 97].

Recueil des données : Lancé en 1994, le consortium a eu pour premières missions de définir les attributs qui devraient être collectés auprès des centres de dépistages qui émaillent le territoire et de mettre en place un système de recueil de l'état de santé des femmes après le dépistage, en lien avec les registres de cancer. Dans la première phase du projet, neuf centres de dépistage ont été intégrés au programme, ils sont situés dans les régions majeures des États-Unis. Ces régions ont été choisies de manière à représenter la population du pays [Sickles 05] afin de diminuer le biais de sélection géographique. En revanche, l'échantillon de la population américaine

que représente cette base de données peut être biaisé si l'on considère le fait qu'elle n'inclut que des femmes ayant fait la démarche de participer au dépistage du cancer du sein, une population qui possède donc au minimum une connaissance des enjeux plus élevée que la moyenne aux États-Unis.

Données publiques : Un extrait de la base de données constituée par le BCSC a été publié en ligne suite à la publication d'une étude traitant de la construction d'un score de risque [Barlow 06]. Cet extrait comporte les données de sept centres de dépistage avec des données collectées entre janvier 1996 et décembre 2002. Lors de leur rendez-vous dans le centre de dépistage pour une mammographie de contrôle, les femmes, âgées de 35 à 84 ans, ont répondu à des questionnaires qui n'avaient pas encore été standardisés. Cette base permet de lier l'indice de densité mammaire évalué par un radiologue selon la méthode BI-RADS (Breast Imaging-Reporting and Data System, [Reston 03]) à un ensemble d'éléments décrivant le profil de chaque femme.

Parmi les 2 392 998 enregistrements de la base de données publique qui correspondent chacun à une mammographie initiale d'une femme et à son profil, les auteurs n'ont pas inclus les femmes qui avaient déjà eu un cancer du sein précédemment ou qui avaient eu une mammographie dans les neuf mois précédents. En revanche toutes les femmes avaient subi une mammographie de contrôle qui datait de cinq ans au maximum. Chaque enregistrement comporte une information sur un éventuel cancer du sein subi dans l'année suivant la mammographie initiale grâce au croisement des données du BCSC et des registres de cancer américains. Cette base de données comporte de nombreuses données manquantes à cause de l'organisation du mode de recueil des données qui a varié au fil des années où l'étude a été menée (les détails sont présentés dans la partie 5.2.1, page 121).

Malgré ces lacunes, elle constitue une des rares sources de données publiquement accessibles sur le cancer du sein en population générale. Elle est d'autant plus rare qu'elle regroupe des facteurs morphologiques et environnementaux sur un large échantillon de femmes quand les autres données disponibles ne concernent généralement que quelques milliers de femmes et regroupent des données biologiques sur les tumeurs. Ces dernières étant inutiles dans la perspective de la construction d'un score de risque pour le cancer du sein pour la population générale.

Outre la confrontation des résultats obtenus par [Barlow 06] et par nous-mêmes grâce à l'utilisation de l'algorithme des plus proches voisins sur cette base publique, nous souhaitons construire un score de risque du cancer du sein pour les femmes françaises. Pour cela, nous avons pu utiliser les données de la plus grande enquête de cohorte française sur le cancer du sein.

3.2.1.2 E3N, une étude de cohorte française

L'Étude Épidémiologique auprès des femmes de la Mutuelle Générale de l'Éducation Nationale (E3N) est une enquête de cohorte prospective* menée à l'INSERM par [Clavel-Chapelon 96, Clavel-Chapelon 97].

Présentation de la cohorte : Après une première phase pilote menée en 1989 dans trois départements (Nord, Pas-de-Calais et Tarn-et-Garonne) auprès de 2 720 femmes, le recrutement a été étendu au plan national auprès d'environ 500 000 femmes adhérentes de la MGEN. Au final, 98 995 femmes, nées entre 1925 et 1950 ont accepté de participer à l'étude en remplissant le premier questionnaire de l'étude et en signant un consentement de participation.

C'est l'équipe 9 du Centre de Recherche en Épidémiologie et Santé des Populations (CESP, INSERM UMR 1018) qui assure la logistique de l'étude à l'IGR. Son objectif est d'étudier la relation entre, d'une part, le mode de vie, les facteurs hormonaux, alimentaires ou reproductifs et, d'autre part, le risque de cancer et de maladies chroniques chez les femmes.

Le recueil des données pour l'étude E3N a été approuvé par la Commission Nationale Informatique et Libertés (CNIL). Les données de la cohorte ne sont pas accessibles publiquement.

Recueil des données : Le recueil des données, toujours en cours, est effectué grâce à des auto-questionnaires que les femmes remplissent à leur domicile et retournent par envoi postal.

Questionnaires principaux Suite à la phase pilote matérialisée par l'envoi du questionnaire zéro (Q0), 10 questionnaires principaux ont été envoyés aux femmes selon la chronologie détaillée dans la figure 3.1, soit tous les 1 à 3 ans environ. Si les questionnaires trois (Q3) et quatre (Q4) ont été envoyés seulement aux femmes ayant répondu au questionnaire précédent (respectivement Q2 et Q3), tous les autres questionnaires (Q1, Q2 et Q5 à Q10) ont été envoyés à toutes les femmes n'ayant pas manifesté leur désir de quitter la cohorte après avoir signé le consentement de participation.

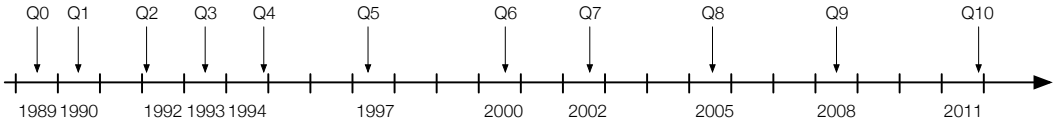

Figure 3.1 : Chronologie d'envoi des auto-questionnaires E3N.

Les questionnaires initiaux (Q0 et Q1) incluaient des questions sur des paramètres anthropométriques à différents âges, des facteurs hormonaux et reproductifs,

l'utilisation de méthodes contraceptives, l'état de santé, les antécédents personnels et familiaux de santé, la consommation de tabac et l'activité physique. Le questionnaire Q2 mettait l'accent sur la prise de traitements hormonaux et les grossesses, tandis que Q3 et Q8 relevaient les consommations alimentaires et Q4 les mesures anthropométriques. En plus de la mise à jour des informations sur l'état de santé et les traitements, les questionnaires suivants ont permis de mettre à jour des informations spécifiques comme le poids, la consommation de tabac ou le statut ménopausique. Certaines informations ont également été recueillies comme la prise de compléments alimentaires, l'autonomie ou la santé bucco-dentaire.

En parallèle de ces questionnaires principaux, des questionnaires ciblés ont été envoyés. Ils ont permis de recueillir des informations plus précises et des documents sur des thèmes particuliers, notamment pour valider médicalement les déclarations des participantes concernant les maladies : par exemple concernant les événements cardio-vasculaires, le cancer du côlon, le suivi mammographique ou encore les cas de diabète.

En cas d'absence de réponse aux questionnaires, des relances ont été effectuées. Pour les questionnaires principaux Q1 à Q9, au moins deux relances ont été réalisées pour un taux de réponse final compris entre 76 et 92 %. Pour le questionnaire Q10, une première relance a été envoyée en octobre 2012.

Suivi de l'état de santé Afin de suivre au plus près l'état de santé des femmes de la cohorte, la MGEN fournit à l'équipe E3N, tous les trimestres depuis 2004, des données partielles sur les remboursements de médicaments dont elle a connaissance ainsi que des données sur le statut vital des participantes qui sont encore adhérentes à la mutuelle. Le décès d'une participante peut ainsi être rapporté par la MGEN, mais aussi par un membre de la famille ou les services postaux. Lorsqu'un tel événement se produit, un courrier est adressé à la mairie de la commune de naissance de la participante afin d'obtenir le lieu et la date de décès. Le Cepi-DC est contacté à intervalles réguliers pour obtenir les causes de décès.

En plus des données de la MGEN, les participantes sont invitées à mettre à jour leur état de santé à chaque questionnaire. Chacune peut déclarer le ou les cancers qu'elle a pu avoir et la date de diagnostic correspondante. Si la participante n'a pas fourni de compte-rendu anatomo-pathologique*, un courrier est adressé à son médecin si son adresse est enregistrée. Dans le cas d'un décès par cancer intervenu entre deux questionnaires et connu par l'acte de décès, les précisions sur l'histoire du cancer pourront être obtenues par l'intermédiaire du médecin si c'est possible. Dans le cas contraire, les femmes sont censurées : le suivi sera considéré comme ayant pris fin au dernier questionnaire précédant le décès.

Pour le cancer du sein, les dernières informations compilées faisaient état d'un niveau de confirmation des cas de cancer par compte-rendu anatomo-pathologique de 92,4 % en 2011.

Représentativité de la cohorte E3N Par le nombre de participantes, la durée de suivi qui excède désormais vingt ans, les taux de réponse enregistrés et la qualité des informations recueillies (voir, par exemple, [Tehard 02, Kesse 02]), la cohorte E3N constitue une source de données unique en France et peu commune au niveau européen.

Si la question de la représentativité de la population française par les participantes à la cohorte E3N peut se poser, il ne faut pas perdre du vue que les études épidémiologiques en général, et la construction de score de risque en particulier, n'ont pas pour objet d'estimer des taux de maladie ou de mortalité au niveau français à partir des données de la cohorte. L'objectif des études épidémiologiques étant la mise en évidence de niveaux de risque différents en fonction des profils et des comportements des participants, la représentativité absolue de la cohorte n'est pas recherchée. Qu'un profil soit plus représenté dans la cohorte que dans la population française ne change pas les liens d'association qui peuvent être mis en évidence entre ce profil et un risque de maladie.

En revanche, la capacité des investigateurs à collecter des données permettant de discriminer des types de profils différents est capitale pour la réalisation des études épidémiologiques. En effet, de mauvais choix à ce niveau conduisent à ne pas pouvoir différencier les profils de participants, ce qui complique la tâche de recherche d'association entre profils et maladie. La diversité des informations collectées et la constance des taux de participation permettent à la cohorte E3N de tendre vers cet objectif.

Indépendamment de leur qualité, l'utilisation de données de santé, dont nous venons de décrire le recueil à travers deux exemples, pose des problèmes dont le déséquilibre des classes malade ou sain, la prise en compte de la confidentialité des données dans leur utilisation et la capacité des patients à utiliser les analyses ou scores de risque pour influer leur risque de maladie.

3.2.2 Déséquilibre des données liées à la santé

Bien que le cancer du sein soit le cancer le plus fréquent chez les femmes, le nombre de nouveaux cas annuels reste faible, relativement à la taille de la population française puisqu'en 2005 on dénombrait, en moyenne standardisée sur l'âge, 101,5 nouveaux cas pour 100 000 femmes. En fouille de données, si on attribue aux femmes touchées par le cancer du sein, la classe positive et aux femmes indemnes, la classe négative, alors la classe positive sera largement minoritaire par rapport à la classe négative. Auparavant, dans les problèmes de fouille de données, on considérait une répartition 90 % contre 10 % comme déséquilibrée et les jeux de données de l'université d'Irvine aux États-Unis [Asuncion 07], très populaires en fouille de données, ont été adaptés dans différentes études pour refléter ce déséquilibre (voir par exemple [García 08]). Par exemple pour le cancer du sein, si on considère une période de temps de 5 ans, ou 10 ans, la classe minoritaire qui regroupe les cas

de cancer du sein en population générale n'est attribuée qu'à 0,5 % des individus, respectivement 1,0 %. On peut donc considérer que la répartition des classes pour le cancer du sein est très déséquilibrée.

Il est donc important de considérer les problèmes que peut poser la fouille de données très déséquilibrées, de la collecte des données à l'évaluation des modèles [Weiss 04].

- Au niveau des données : un déséquilibre de la répartition des classes de données peut avoir deux origines. Les données labellisées de la classe minoritaire peuvent être peu nombreuses d'un point de vue relatif à la classe majoritaire, mais également d'un point de vue absolu. Cette rareté absolue de la classe minoritaire est la plus problématique puisqu'il peut être très difficile de collecter assez de données pour permettre l'utilisation d'un algorithme de fouille. La rareté relative pose moins de problèmes puisqu'elle pourra être compensée par la collecte d'un nombre élevé de données pour augmenter la quantité de données de la classe labellisée positive. Par exemple, dans le cas du cancer du sein, la rareté absolue est compensée par l'existence de structures organisées de recueil de données que sont les enquêtes de cohorte. Bien que peu nombreuses, les grosses cohortes permettent de regrouper suffisamment d'individus et de les suivre sur une période suffisamment longue pour que le nombre absolu de cas de cancer atteigne un minimum nécessaire à la validité statistique des risques estimés.
- Au niveau des algorithmes : d'une part, l'utilisation d'algorithmes habituels de fouille de données qui tirent parti de la stratégie *diviser pour conquérir* peut poser problème avec des données déséquilibrées. En effet, ce type de stratégie conduit à diviser l'espace de recherche (la population) en sous-espaces (les sous-groupes de populations) de tailles de plus en plus faibles pour trouver les motifs ou les régularités recherchées. Cette fragmentation conduit à diminuer encore le nombre absolu de cas de la classe minoritaire dans chaque sous-espace de recherche, rendant plus problématique le déséquilibre des données. D'autre part, l'effet du bruit dans les données peut être démultiplié en cas de déséquilibre des données. Il faudra en effet bien moins de données parasites pour bruiter une classe minoritaire, rendant d'autant plus difficile l'apprentissage des algorithmes de fouille.
- Au niveau de l'évaluation : le déséquilibre des données peut également influer tout le processus de fouille par l'intermédiaire des méthodes d'évaluation qui y jouent un rôle essentiel. Par exemple, dans un problème de prédiction de classes binaires où 90 % des données appartiennent à la classe négative et 10 % à la classe positive, si l'algorithme conduit à classer tous les nouveaux enregistrements comme appartenant à la classe majoritaire, y compris ceux qui devraient être classés comme appartenant à la classe minoritaire, le taux des enregistrements correctement classés sera élevé bien que le classifieur ne fonctionne pas du tout. Cela a été vérifié de manière empirique par [Weiss 03]. Des méthodes d'évaluation qui permettent de prendre en compte les coûts de

mauvaise classification ou qui n'avantagent pas la classe majoritaire, doivent être utilisées. C'est le cas de l'aire sous la courbe ROC dont la construction est détaillée en partie 2.3.2.1, page 50.

3.2.3 Prise en compte de la confidentialité

Si la confidentialité des données doit être respectée dans tout processus de fouille, le législateur a mis l'accent sur la protection des données personnelles, et plus encore des données de santé. En France, la Commission Nationale de l'Informatique et des Libertés (CNIL), l'autorité de contrôle française en matière de protection des données personnelles, veille à ce que soient prises les « dispositions nécessaires pour assurer la sécurité des données enregistrées et empêcher qu'elles ne soient divulguées ou utilisées à des fins détournées surtout s'il s'agit d'informations couvertes par le secret médical. » En outre, elle « préconise l'adoption de mesures de sécurité physique et logique qui doivent être adaptées » en fonction du contexte d'utilisation.

Cette problématique doit donc être prise en compte dans le processus de construction d'un score de risque. En effet, la modélisation d'un risque de maladie suppose l'utilisation de données couvertes par le secret médical : d'une part la description du profil des personnes en fonction de caractéristiques diverses comme l'âge, les caractères morphologiques ou les antécédents de traitements médicaux par exemple et, d'autre part, les antécédents de maladies qui sont à la base des méthodes de prédiction de risque puisqu'ils permettent de construire des modèles de risque par la recherche de motifs liant les profils des personnes et les maladies qu'elles ont subies.

Dans le cas de la cohorte E3N par exemple, les femmes qui y participent ont signé un consentement lors du remplissage du premier questionnaire qui vaut pour l'utilisation de leurs données dans le cadre de la recherche médicale. Afin d'assurer la confidentialité des données transmises par ces femmes, chacune d'entre elles est identifiée dans les bases de données par un numéro d'anonymat qui permet de réaliser les associations entre les données lors de la manipulation des différentes bases. La correspondance avec l'identification par nom, prénom et adresse pouvant être réalisée seulement par les personnes habilitées conformément aux recommandations de la [CNIL 12].

L'anonymisation des données permet de garantir la confidentialité des données, y compris dans le cadre de l'utilisation de l'algorithme des plus proches voisins dont le principe de base implique de disposer de la base de données pour réaliser une prédiction du risque de maladie. Pour limiter la diffusion de données, même anonymisées, un second niveau de protection est offert par le précalcul des niveaux de risque qui seront inclus dans un outil d'évaluation du risque de la maladie à destination de l'utilisateur final. Tous les niveaux de risque qui peuvent potentiellement être rencontrés sont précalculés et seule cette base des scores précalculés est intégrée dans l'outil mis à disposition de l'utilisateur final.

3.2.4 Possibilité d'action par rapport au niveau de risque

Un autre problème, lié à l'utilisation de données de santé dans le but de prédire un risque, est la capacité des acteurs de la prévention, et l'utilisateur final en premier lieu qu'il soit patient ou médecin, à pouvoir modifier le comportement d'une personne pour influer sur le risque prédit de maladie.

Selon [Silberschatz 96], l'utilisabilité est une mesure majeure de l'intérêt d'un motif –l'association d'un type de profil et d'un risque élevé de maladie– mis en évidence par un algorithme de fouille de données. Pour un score de risque de maladie, cela signifie que le résultat de l'évaluation d'un risque doit permettre au médecin ou au patient de décider d'une ou plusieurs actions concrètes pour diminuer le risque. Une partie de ces actions est le fait de pouvoir modifier les facteurs de risque de la maladie pour diminuer le niveau de risque. Si pour certaines maladies, il existe des facteurs de risque identifiés facilement modifiables (tabac pour le cancer du poumon, nutrition et activité physique pour les maladies cardio-vasculaires), c'est moins le cas pour le cancer du sein. En effet peu des facteurs de risque, présentés en partie 1.2.3, page 17, peuvent être modifiés : l'âge de la femme est fixé, les facteurs hormonaux tels que l'âge aux premières règles, à la ménopause ne peuvent être contrôlés, pas plus que les antécédents qui révèlent la part génétique de la maladie ou la densité mammaire. En revanche, la prise de traitements hormonaux peut être maitrisée. De même que les facteurs liés aux habitudes de vie comme l'activité physique ou la nutrition, même s'ils n'ont qu'un faible impact sur le niveau de risque.

Pour les maladies comme le cancer du sein, si l'action qui découle de la mesure du niveau de risque ne peut pas (ou relativement peu) se faire en modifiant un comportement, elle peut en revanche se faire au niveau de la modification de la surveillance. Par exemple, la fréquence du dépistage du cancer du sein pourra être adaptée au niveau de risque de la femme. C'est une partie du projet de clinique du risque proposé par l'IGR (voir page 9).

Cette possibilité d'action sur les facteurs de risque fait partie des connaissances qui doivent être apportées par les experts du domaine, dans le cadre du processus de conception de score de risque que nous proposons.

3.3 Proposition d'un processus basé sur le modèle de processus CRISP-DM

Pour répondre aux objectifs fixés dans le domaine de la santé publique, en prenant en compte les contraintes que nous avons détaillées, et pallier les faiblesses des modes de construction des scores de risque existants, nous présentons dans cette partie un processus de création et d'utilisation de score de risque dans le domaine médical qui est basé sur le modèle de processus CRISP (pour « CRoss Industry Standard Process for Data Mining »). Dans cette optique, nous présentons d'abord le modèle de processus CRISP-DM puis les parties prenantes des processus que nous décrivons par la suite et le type de connaissances qu'ils apportent. Enfin, nous présentons deux processus, le premier pour créer un score de risque adapté au contexte grâce à la connaissance des parties prenantes et le second qui modélise l'utilisation du score de risque.

3

3.3.1 Présentation du modèle de processus CRISP-DM

Le modèle de processus CRISP-DM a été développé à partir de 1996 dans l'industrie afin de proposer une approche standardisée d'un problème de fouille de données qui soit librement disponible et non-propriétaire [1]. Selon ses concepteurs, il est basé « sur la pratique et l'expérience du monde réel ». Longtemps resté à l'état de travail en cours, il a cependant été utilisé par plusieurs grandes entreprises pour conduire des projets, avant d'être publié en version finale 1.0 en 2000.

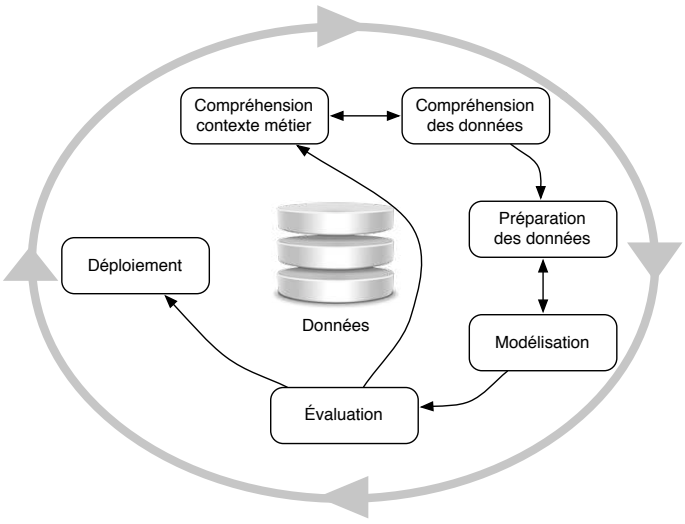

Figure 3.2 : Phases du modèle de référence CRISP-DM selon [Chapman 00].

1. En partie financé par le projet européen ESPRIT numéro 24 959 qui regroupait Daimler-Chrysler AG, SPSS, NCR et OHRA.

Le modèle de processus CRISP-DM peut être décrit à plusieurs niveaux : deux niveaux génériques et deux niveaux spécialisés. Au premier niveau générique, six phases sont distinguées (voir figure 3.2) et chacune de ces phases regroupe plusieurs tâches génériques qui constituent le deuxième niveau. L'objectif est de couvrir le plus grand nombre de situations tout en gardant un processus qui soit complet et stable. À partir du troisième niveau, les tâches décrites sont spécialisées en fonction du domaine d'utilisation et ne sont plus génériques. Le quatrième niveau est constitué de plusieurs processus appliqués au problème qui décrivent chaque tâche du troisième niveau : ces tâches et ces processus doivent être adaptés par l'utilisateur au cas concret. Par exemple, on trouve dans le deuxième niveau de la phase de modélisation, une tâche *construction du modèle*. Celle-ci devra être adaptée au problème que l'on cherche à résoudre par la réalisation d'actions concrètes comme le paramétrage du modèle choisi ou la récupération physique des résultats obtenus.

L'enchaînement des phases et des tâches n'est pas figé et s'adapte à la vie du projet de fouille de données. En particulier, l'enchaînement des phases, symbolisé par des flèches sur la figure 3.2, montre les dépendances les plus importantes et les plus fréquentes sans qu'elles constituent un chemin à suivre de manière impérative. Il faut également noter que le modèle de processus CRISP-DM est itératif en fonction des évaluations qui sont réalisées à chaque phase du processus. Le déroulement logique, mais pas impératif, car des boucles sont possibles, est décrit ci-dessous.

- *Compréhension du métier* : La phase initiale du processus est dédiée à la compréhension du problème métier, sa transformation en un problème de fouille de données et l'élaboration d'un plan préliminaire pour sa résolution. Cette phase doit permettre de comprendre des objectifs métiers parfois contradictoires. Pour la construction d'un score permettant de mesurer le risque de maladie, les objectifs contradictoires peuvent être la nécessité d'une haute performance dans la détection des profils à risque et la volonté d'adapter le score au contexte d'utilisation qui implique l'utilisation de facteurs de risque avec un pouvoir prédictif moins fort. Les objectifs métiers, les objectifs de fouille de données et le plan du projet peuvent être produits en sortie de cette phase. La majorité des chapitres 1 et 2 de ce manuscrit est consacrée à la compréhension du métier avec le cancer du sein en exemple.
- *Compréhension des données* : La compréhension des données débute par la collecte des données et la familiarisation avec celles-ci dans le but de détecter d'éventuels problèmes de qualité des données et de formuler de premières hypothèses quant à la résolution du problème. Un rapport sur le mode de recueil des données, leurs distributions et leurs valeurs manquantes peut être produit en sortie de cette phase. Pour le cancer du sein, cette phase est traitée en partie 1.2.3, page 17, par l'analyse des facteurs de risque de la maladie et en partie 5.2, page 121, par la présentation des données à disposition et l'analyse de leurs spécificités.
- *Préparation des données* : La préparation des données regroupe les activités de construction d'un jeu de données utilisable par les outils de modélisation à

partir des données brutes collectées. Cette phase inclut le nettoyage des données, leur transformation et pourra être réalisée de multiples fois. Un jeu de données et un rapport contenant les actions conduites pour les rendre modélisables (par exemple concernant la gestion des valeurs manquantes et la discrétisation des valeurs) peuvent être produits en sortie de cette phase. Les choix réalisés dans le cadre de la construction d'un score de risque de cancer du sein sont présentés en partie 5.2, page 121.

- *Modélisation* : Cette phase inclut la sélection de la technique de modélisation, la définition des paramètres qui y sont relatifs et son application aux données. Les exigences sur les données peuvent être différentes selon la technique de modélisation retenue, un retour à la phase de préparation des données peut donc être nécessaire en fonction du choix effectué. Une description de l'algorithme utilisé et les résultats de la modélisation des données, exprimés sous la forme de mesures de performances notamment, peuvent être produits en sortie de cette phase. Pour notre application au cancer du sein, la technique de modélisation et les résultats qui en découlent sont présentés en partie 5.1, page 109, et partie 5.3, page 132, grâce à l'utilisation de l'architecture présentée au chapitre 4.
- *Évaluation* : À cette étape, un ou plusieurs modèles ont été retenus du point de vue de l'adéquation aux données. L'évaluation permet de passer en revue les précédentes étapes et de poser la question de la validité de la réponse apportée au problème métier. Si un ou plusieurs aspects du problème métier n'ont pas trouvé de réponse, un retour à l'étape de compréhension du problème métier est nécessaire. Le processus de vérification de l'adéquation de la réponse apportée au problème métier, le résultat de cette adéquation et les perspectives d'améliorations peuvent être consignés dans un document produit en sortie de cette phase. Cette phase d'évaluation des résultats obtenus correspond à la discussion des résultats obtenus au regard de la littérature et des objectifs initiaux, en partie 5.3.1.3, page 134, et en partie 5.3.2.3, page 140.
- *Déploiement* : Que le déploiement de la solution apportée corresponde à la simple écriture d'un rapport ou, au contraire, à la mise en place d'un système informatique complexe basé sur les résultats du processus de fouille, la clef de la réussite d'une phase de déploiement est une bonne communication entre l'analyste qui a déroulé le cycle et le client qui prend possession des résultats de la fouille de données. En sortie peuvent être produits un plan de déploiement, un plan de maintenance, un rapport final sur le déroulement de tout le processus et une revue du projet sous un angle critique, de retour d'expérience et de documentation. Pour le score de risque de cancer du sein, les besoins et un prototype d'outil déployé sont détaillés en partie 3.4.6, page 82. Ce manuscrit de thèse constitue une partie de la revue du projet sous les angles énoncés.

À partir de ce modèle de processus, nous proposons un processus dédié à la conception et à l'utilisation d'un score de risque dans le domaine médical que nous présentons dans la partie suivante 3.3.2 et dont nous détaillons l'utilisation pour le

cas du cancer du sein dans la partie 3.4, page 79.

3.3.2 Utilisation de la connaissance dans le processus de conception du score de risque

Afin de comprendre l'action des différents acteurs du processus de conception d'un modèle de risque, nous présentons dans un premier temps les parties prenantes et le type de connaissances qu'ils apportent au cours du processus de conception du score de risque. Nous présentons ensuite un processus de conception de modèle de risque et celui permettant son utilisation en nous basant sur le modèle de processus CRISP-DM présenté en partie 3.3.1.

3.3.2.1 Identification des parties prenantes et leurs types de connaissance

La conception d'un score de risque en santé, grâce à une méthode de fouille de données, suppose l'intervention de différentes personnes dont nous explicitons le rôle.

Le score de risque est conçu pour des *utilisateurs* qui peuvent être de différentes natures. L'utilisateur peut être une personne du grand public qui a peu ou pas de notions des facteurs de risque de la maladie pour laquelle il utilise le score. L'utilisateur peut également faire partie de la communauté médicale avec un niveau de spécialisation plus ou moins élevé, depuis le médecin généraliste qui utilise le score pour réaliser de la prévention en population générale jusqu'au spécialiste d'une maladie qui utilise un score construit sur une population spécifique pour évaluer le risque d'un patient. Dans ce dernier cas, on parle alors d'utilisateur expert.

Si l'utilisateur n'apporte pas de connaissance proprement dite au processus de conception d'un modèle de risque, il possède en revanche certaines connaissances empiriques et certains a priori sur le risque de maladie. Certaines de ces connaissances sont valides scientifiquement tandis que d'autres peuvent ne plus l'être, notamment en raison de l'avancée de la recherche scientifique. Ces a priori doivent être pris en compte lors du processus de conception du score de risque comme moyen d'en faciliter l'acceptation et c'est notamment le rôle de l'analyste du besoin de les synthétiser en tant que besoins des utilisateurs.

Dans notre proposition, nous attribuons à l'*analyste* du besoin la connaissance des besoins des utilisateurs en amont du processus de conception du score et la possibilité d'intervenir dans le processus pour s'assurer du respect des contraintes liées à ces besoins utilisateurs. Son rôle est de synthétiser les besoins des utilisateurs que ce soit sur le plan des attributs disponibles pour évaluer le risque et sur celui des a priori des utilisateurs sur le risque.

Les processus, que nous présentons dans la suite de cette partie, nécessitent l'intervention d'un expert du domaine. Nous avons choisi de faire appel à un *épidémiologiste* qui doit connaître à la fois les principaux mécanismes biologiques de la maladie étudiée et les facteurs de risque qui influencent le risque de maladie. L'épidémiologiste apporte sa connaissance des données, depuis leur recueil jusqu'à leur

distribution et leur impact sur le niveau de risque mesuré. Il permet d'assurer la cohérence des choix effectués au long du processus avec la littérature épidémiologique.

Enfin, pour s'assurer de la maîtrise de l'utilisation d'un algorithme de fouille de données, un *expert en fouille de données* est nécessaire lors de l'utilisation d'algorithmes qui ne sont pas utilisés habituellement pour construire un score de risque. Il apporte sa connaissance des algorithmes : leur mode de fonctionnement, leur complexité ou leur configuration. En revanche, lors de l'utilisation de systèmes informatiques intégrant des algorithmes déjà mis en œuvre, il n'est plus nécessaire de faire intervenir un expert de fouille de données dont la connaissance est principalement nécessaire à la conception du processus et à l'ajout de nouveaux algorithmes.

Les quatre rôles que nous venons de décrire interviennent dans les processus de conception d'un modèle de risque et d'obtention du niveau de risque que nous décrivons dans la partie suivante.

3.3.2.2 Conception d'un modèle de risque

Le processus de conception d'un modèle de risque correspond aux phases du modèle de processus CRISP-DM de préparation des données, de génération des modèles de risque et de choix du modèle de risque, mais elle débute par la constitution d'une liste d'attributs extraits d'une base de données que nous décrivons en premier.

Conception d'une liste d'attributs : Les bases de données, qui permettent de stocker les informations relatives au suivi des cohortes* utilisées par les épidémiologistes, contiennent plusieurs types de données (brutes, nettoyées, générées, détail page 125) et il n'est pas envisageable de tester les milliers d'attributs de ces bases de données dans le cadre d'un travail de conception d'un score de risque, notamment pour des raisons de puissance de calcul. Nous avons donc fait le choix d'imposer une phase de limitation de l'espace de recherche au cours de laquelle, l'expert en fouille de données en charge de la construction du score, l'épidémiologiste et l'analyste doivent limiter le nombre d'attributs qui seront testés.

Une première phase de sélection des attributs est réalisée par l'expert en fouille de données suite à la phase de compréhension du problème métier. Cette phase consiste à sélectionner les attributs contenus dans une base de données épidémiologique en fonction d'un possible lien avec la maladie dont on cherche à prédire le risque. Cette liste est réalisée en fonction de la littérature du domaine et doit être large.

Une phase de filtrage est ensuite réalisée par l'épidémiologiste en fonction de sa connaissance précise de la maladie. À ce stade, certains attributs, qui influent le risque, mais de manière marginale, peuvent être éliminés.

Enfin, l'analyste, en fonction de sa connaissance de besoins des utilisateurs, notamment traduits en termes de coût d'acquisition des attributs (voir partie 3.1.2.2), décide de la validation, ou non, de la liste filtrée qui a été produite. En cas de désaccord, des discussions entre les trois acteurs de ce processus sont réalisées pour aboutir à la production d'une liste dont les attributs pourront être inclus dans le

score de risque en fonction des performances mesurées dans l'étape de modélisation.

Préparation des données : La phase de préparation des données est une occasion d'intégrer de la connaissance au processus de fouille [Brisson 08, Brisson 09]. En utilisant différentes règles de décisions discutées par les acteurs du processus, l'expert en fouille de données prépare un jeu de données qui inclut les attributs sélectionnés à l'étape précédente. Les règles de décision concernent la manière dont sont discrétisées les données (si l'algorithme utilisé le nécessite), la manière dont les données manquantes sont gérées (suppression ou imputation par exemple) et la manière dont certains attributs ont pu être reconstruits à partir de données obtenues de différentes sources. Par exemple, la reconstruction de la prise d'un médicament par une personne peut être effectuée en fonction de questionnaires déclaratifs à l'aide du croisement de fichiers issus du système de santé national, ce qui impose de prendre des décisions sur la manière de procéder, notamment concernant les données à privilégier, par exemple quand les qualités de données sont différentes.

Une fois le jeu préparé, un document est rédigé pour décrire le jeu de données, du point de vue la distribution des valeurs par attributs et des règles de décision utilisées afin de passer une étape de validation réalisée par l'épidémiologiste. Celui-ci vérifie la cohérence des données en confrontant les informations du document à sa connaissance de la maladie. En cas de désaccord ou d'aberration dans les données, l'étape de préparation des données est de nouveau réalisée avant l'étape de modélisation.

Choix du modèle de risque : À partir du jeu de données validé, des modèles de risque sont conçus. Ces modèles sont basés sur un algorithme et sa configuration d'une part, et sur un sous-ensemble des attributs sélectionnés d'autre part. Plusieurs dizaines à plusieurs milliers de modèles sont générés sous la responsabilité de l'expert en fouille de données si l'on prend en compte les différentes combinaisons d'algorithmes testés, les configurations associées, les tailles de combinaisons d'attributs et les différentes répartitions de données dans les jeux d'apprentissage et de validation.

Une fois les modèles générés, leur performance est testée, en utilisant la calibration et la discrimination, toujours sous la responsabilité de l'expert ou fouille de données.

Les résultats de ces mesures de performance sont fournis à la fois à l'épidémiologiste et à l'analyste qui doivent choisir un modèle ou valider le modèle choisi par l'autre. Ce choix est réalisé par des experts qui utilisent leur connaissance de la maladie (pour l'épidémiologiste) et leur connaissance des besoins des utilisateurs (pour l'analyste). Il doit prendre en compte la performance du score à détecter les personnes à risque et, de nouveau, le contexte d'utilisation du score de risque dont le type d'utilisateur du score. En cas de performance équivalente pour différentes combinaisons d'attributs, la simplicité du score peut entrer en ligne de compte. Cette simplicité peut être exprimée par la réduction du nombre d'attributs utilisés, ou

au contraire, par la réalisation d'un choix qui met en valeur de certains attributs considérés comme prépondérants dans le risque de maladie (et reconnu comme tel par les autorités de santé par exemple).

À partir du modèle sélectionné et validé par l'épidémiologiste et l'analyste, une table de référence (base de données permettant d'associer un profil à son niveau de risque précalculé) est créée en vue de la phase de déploiement du score de risque auprès des utilisateurs.

3.3.2.3 Obtention d'un niveau de risque

Le processus d'obtention d'un niveau de risque correspond à la description des fonctions de l'outil qui, après déploiement du modèle de risque dans un système informatique (phase de déploiement dans le modèle de processus CRISP-DM), doit permettre à l'utilisateur d'obtenir un niveau de risque pour le profil d'une personne.

La première phase consiste en la saisie d'un profil de personne en fonction des attributs retenus lors de la phase de choix du modèle de risque et qui sont stockées dans la table de référence. Une femme dont on souhaite déterminer le risque de cancer du sein doit renseigner son profil, par exemple en termes d'âge ou de nombre d'antécédents de cancer du sein.

Lors de la seconde phase, l'évaluation du profil est réalisée en interrogeant la table de référence et un niveau de risque est fourni. Par exemple pour le cancer du sein, le système informatique fait correspondre le profil renseigné à un niveau de risque contenu dans une table de référence.

Les objectifs fixés en matière d'utilisation et d'appropriation pour le système informatique en charge de ce sous-processus sont détaillés à la fin de ce chapitre, page 82, avec une proposition d'interface répondant à ces objectifs.

3.4 APPLICATION DE NOTRE PROCESSUS SUR L'EXEMPLE DU CANCER DU SEIN

Dans cette partie, nous proposons d'illustrer le déroulement du processus théorique proposé sur le cas du cancer du sein en utilisant une base de données regroupant des informations sur des femmes françaises. Les étapes du modèle de processus CRISP-DM sont reprises selon le processus proposé en partie 3.3.2, page 75 : elles couvrent la compréhension du contexte métier, la préparation des données par rapport au contexte d'utilisation du score, la modélisation des données avec une méthode compréhensible et le choix d'une combinaison adaptée, l'évaluation des résultats obtenus et la phase de déploiement du score.

3.4.1 Compréhension contexte métier

Diminuer l'incidence du cancer du sein dans la population française, et la mortalité qui en résulte, nécessite de combattre la maladie sur le front des traitements et de la prévention. Une manière d'augmenter l'efficacité de la prévention est, au-delà des programmes nationaux de dépistage en population générale qui restent indispensables, d'informer et de dépister les femmes les plus à risque : c'est la prévention sélective. Cette détection peut se faire lors de différents événements de la vie médicale de la femme comme chez un médecin généraliste ou chez un radiologue à l'occasion de la réalisation d'une mammographie.

Cette détection bénéficierait d'un score de risque qui permette de cibler les femmes à risque. Les scores existants sont en effet très peu utilisés et nous émettons l'hypothèse que c'est en partie à cause des méthodes de modélisation du risque difficilement compréhensibles par les utilisateurs (médecins ou patientes) et en partie à cause de l'inadaptation de ces outils à une utilisation rapide et facile.

Traduites en un problème de fouille de données, ces hypothèses nous conduisent à vouloir, d'une part, construire un score de risque basé sur une méthode compréhensible comme l'algorithme des plus proches voisins dont nous souhaitons confirmer l'efficacité face aux méthodes de modélisation habituellement utilisées. Les facteurs de risque utilisés doivent être adaptés au contexte d'utilisation défini, tous les attributs ne sont en effet pas disponibles au même coût (en termes de temps, de finance, d'émotion) selon l'endroit où est utilisé le score de risque. Cette volonté d'adaptation au contexte d'utilisation est un objectif contradictoire à l'objectif de performance. D'autre part, la phase de déploiement du score de risque dans un outil mis à la disposition des médecins et des patientes ne doit pas être négligée pour assurer un confort d'utilisation qui favorise sa diffusion.

Le contexte d'utilisation choisi est une clinique du risque dans laquelle un besoin d'information des patientes a été identifié. Les questions posées pour mesurer le risque doivent donc être simples et ne pas nécessiter d'avoir recours à un examen médical ou une recherche particulière.

3.4.2 Compréhension des données

Construire un score de risque à destination des femmes françaises nécessite l'utilisation de données françaises pour modéliser le risque. Nous avons utilisé la cohorte E3N maintenue par l'INSERM qui regroupe 100 000 femmes interrogées tous les 2 à 3 ans depuis 1990 sur leur mode de vie, leur alimentation, leurs traitements médicaux et leur état de santé en général.

Ces données ont plusieurs caractéristiques notables : elles sont très déséquilibrées au sens où une faible proportion des femmes est atteinte du cancer du sein au cours de la fenêtre de temps pour laquelle le score de risque est construit. Cela implique de choisir une méthode de modélisation du risque qui supporte ce déséquilibre. De même, les méthodes de mesure des performances de la capacité à détecter les profils à risque doivent permettre la prise en compte de ce déséquilibre : nous avons choisi de mesurer la capacité de discrimination* des scores de risque produits grâce à l'aire sous la courbe ROC* et au risque relatif observé*. La calibration* du score est déterminée en mesurant le rapport du nombre estimé de cas de cancer sur le nombre observé de cas de cancer dans les données.

La confidentialité des données de santé doit également être prise en compte, notamment en assurant l'anonymat des femmes interrogées. Enfin, la capacité d'action de la patiente et du médecin quant à la possibilité de modifier le risque prédit doit être prise en compte dans la construction du score. En effet, tous les facteurs de risque du cancer du sein ne sont pas modifiables : l'âge de la patiente, le nombre d'antécédents familiaux de cancer du sein ou l'âge aux premières règles sont des facteurs reconnus comme étant importants dans le risque de cancer du sein, mais ne peuvent pas être modifiés afin de diminuer le risque d'une patiente.

À partir de ces contraintes, une large liste d'attributs, qui représentent les facteurs de risque, est constituée en fonction de la littérature : si un impact a été mesuré dans la littérature et si l'attribut est disponible dans la cohorte E3N, l'attribut est ajouté à la liste. Cette liste contient 16 attributs.

Conformément au processus que nous proposons, la liste est soumise à un filtrage par un expert en épidémiologie qui possède la connaissance de la base de données E3N et la connaissance de la littérature scientifique. La liste est soumise à la validation d'un analyste qui a la connaissance des besoins exprimés en termes de disponibilité des attributs dans le contexte d'utilisation envisagé. Des attributs comme l'indice de masse corporelle ou le nombre d'avortements ne sont pas retenus à cause de leur faible impact connu sur le risque. À l'inverse, la mesure de l'activité physique est conservée malgré son impact moyen sur le risque, car il s'agit d'un facteur modifiable qui permet de faire évoluer le niveau de risque d'une patiente à la baisse. La liste filtrée et validée contient 9 attributs.

3.4.3 Préparation des données

Suivant le processus proposé, des règles de décisions sont définies pour des aspects importants comme la gestion des valeurs manquantes ou la discrétisation de certains attributs à valeurs continues.

Deux règles ont principalement été appliquées sur le jeu de données, construit à partir de la cohorte E3N, pour gérer les valeurs manquantes au temps choisi comme base, le questionnaire 5 en 1997. En cas de valeur manquante à ce questionnaire, la valeur est recherchée dans un autre questionnaire pour les attributs dont les valeurs n'ont pas été modifiées d'un questionnaire à l'autre (âge aux premières règles par exemple). Si aucune valeur n'est disponible ou si la valeur de l'attribut peut avoir changé (nombre d'antécédents du cancer du sein par exemple), la médiane est imputée afin de combler la valeur manquante.

La discrétisation est réalisée en tenant compte de la distribution des modalités, en accord avec les pratiques habituelles identifiées dans la littérature et en fonction du contexte d'utilisation de l'outil déployé qui permettra de mesurer le risque.

Le score de risque est conçu pour un horizon de dix ans, l'attribut cible vaut 0 si aucun cancer du sein n'est déclaré entre la date de remplissage du questionnaire Q5 en 1997 et la fin de la période de dix ans qui suit. Sinon, l'attribut cible vaut 1.

La préparation des données est validée par un épidémiologiste sur la base d'un document décrivant les décisions prises et la distribution des données après préparation du jeu de données.

3.4.4 Modélisation

Nous choisissons d'utiliser l'algorithme des plus proches voisins pour modéliser le risque. Pour chaque profil de femme à évaluer, une distance est calculée avec toutes les femmes de la base d'apprentissage. Les k femmes les plus proches sont recrutées dans le voisinage. La prévalence, qui correspond au nombre de femmes atteintes par le cancer du sein sur le nombre total de femmes du voisinage, est calculée : elle correspond au niveau de risque de la femme évaluée.

Parmi les attributs retenus, les performances des combinaisons comprenant 2 à 5 attributs sont mesurées en termes de discrimination et de calibration pour différentes tailles de voisinages et différents jeux de validation croisée. Les résultats, classés par ordre de discrimination (la taille de voisinage retenue est celle qui donne le meilleur résultat), sont présentés à un épidémiologiste et un analyste chargé de répondre aux besoins clients.

Après filtrage de la liste en fonction du critère de la présence impérative de l'âge de la femme, une combinaison incluant l'âge de la femme, le nombre de maladies bénignes du sein, l'âge à la ménopause et le nombre d'antécédents de cancer du sein au premier degré est choisie. Sa discrimination, calculée sous la forme d'une aire sous la courbe ROC, est mesurée à 0,63 pour les tailles de voisinage comprises entre 2 000 et 3 000 voisins. Le risque relatif observé est de 4,8 ce qui signifie qu'il y a 4,8

fois plus de malades du cancer du sein dans le premier décile de scores du jeu de validation que dans le dernier décile.

En ce qui concerne la calibration, le rapport du nombre estimé de cas de cancer du sein sur le nombre observé de cas de cancer du sein est de 0,99. Cela signifie le nombre de cas de cancer du sein prédit est très proche du nombre observé de cas de cancer du sein dans le jeu de validation.

3.4.5 Évaluation

Au regard des objectifs de départ, on évalue la combinaison d'attributs choisie.

Le score est adapté au contexte par le faible coût d'obtention des attributs retenus pour constituer le score qui ne nécessite pas d'examens médicaux et un faible effort de mémoire de la part de la patiente.

Tout en respectant la contrainte d'adaptation au contexte et la contrainte d'utilisation d'une méthode de modélisation compréhensible, les performances mesurées sont comparables aux scores de risque du cancer du sein disponibles dans la littérature dont les performances varient de 0,58 [Gail 89] à 0,64 [Barlow 06] alors que contrairement aux résultats de Barlow, nous n'avons pas utilisé un attribut fortement prédictif comme la densité mammaire [Rockhill 01].

Les objectifs de départ sont donc atteints et la solution peut être déployée.

3.4.6 Déploiement

Afin de faciliter l'utilisation du score de risque, nous avons choisi de construire un outil sous la forme d'une application web. À partir des observations effectuées sur les outils d'évaluation du risque (voir partie 2.2.3, page 42), plusieurs prototypes ont été réalisés. Les interfaces graphiques que nous présentons en figures 3.3 et 3.4, sont les maquettes de la version en cours de développement qui s'inspirent des prototypes. L'affichage est généré en HTML (« HyperText Markup Language », le langage utilisé pour toutes les pages web sur internet) grâce à un moteur PHP (« PHP :Hypertext PreprocessorHypertext Preprocessor ») à partir de données précalculées lors de l'étape de modélisation et contenues dans une base MySQL (un système de gestion de bases de données). Les mises à jour de la page web sont réalisées à l'aide du langage de programmation Javascript.

Le graphique 3.3 présente un extrait des trois écrans qui permettent d'expliquer la méthode de modélisation du risque.

Le graphique 3.4, présente l'interface de l'outil de calcul du score de risque. Nous avons choisi d'afficher l'âge de la femme en abscisse et le niveau de risque calculé en ordonnée. La courbe représente le risque moyen par âge qui permet de placer une référence afin d'apprécier le niveau de risque qui sera calculé de manière individuelle. Le risque de la femme qui utilise le logiciel est calculé à partir d'une table de référence anonyme contenant les scores calculés par l'algorithme en fonction des éléments renseignés pour les différents attributs sous le graphique. Il est affiché

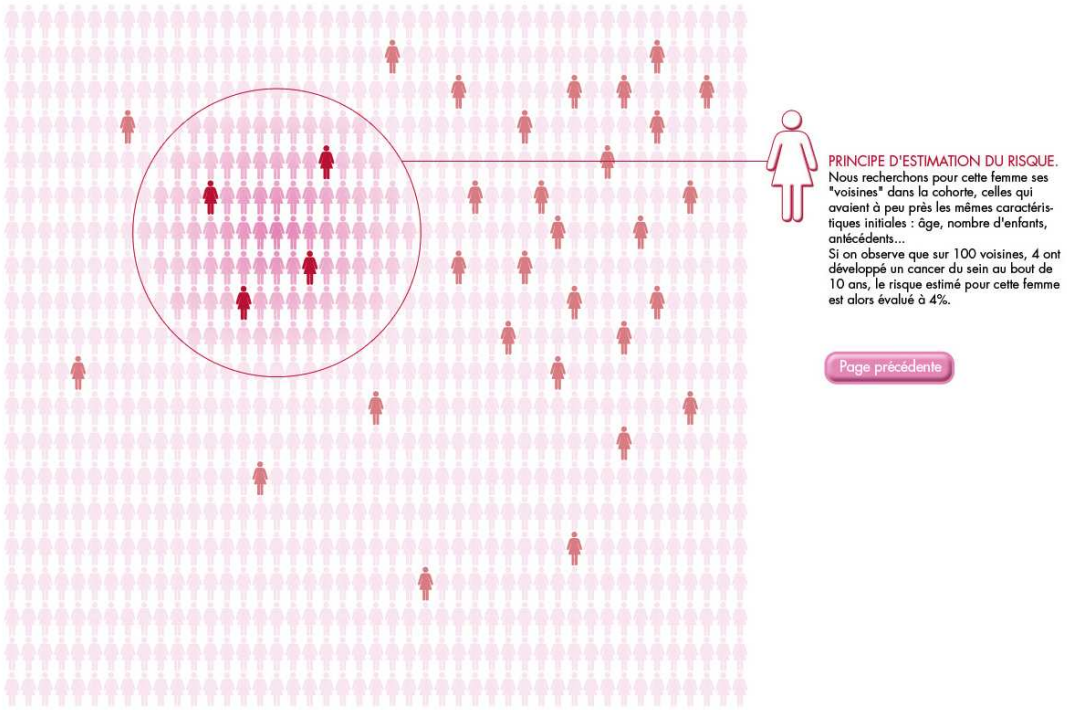

Figure 3.3 : Maquette d'application web pour expliquer le concept de la méthode de modélisation utilisée.

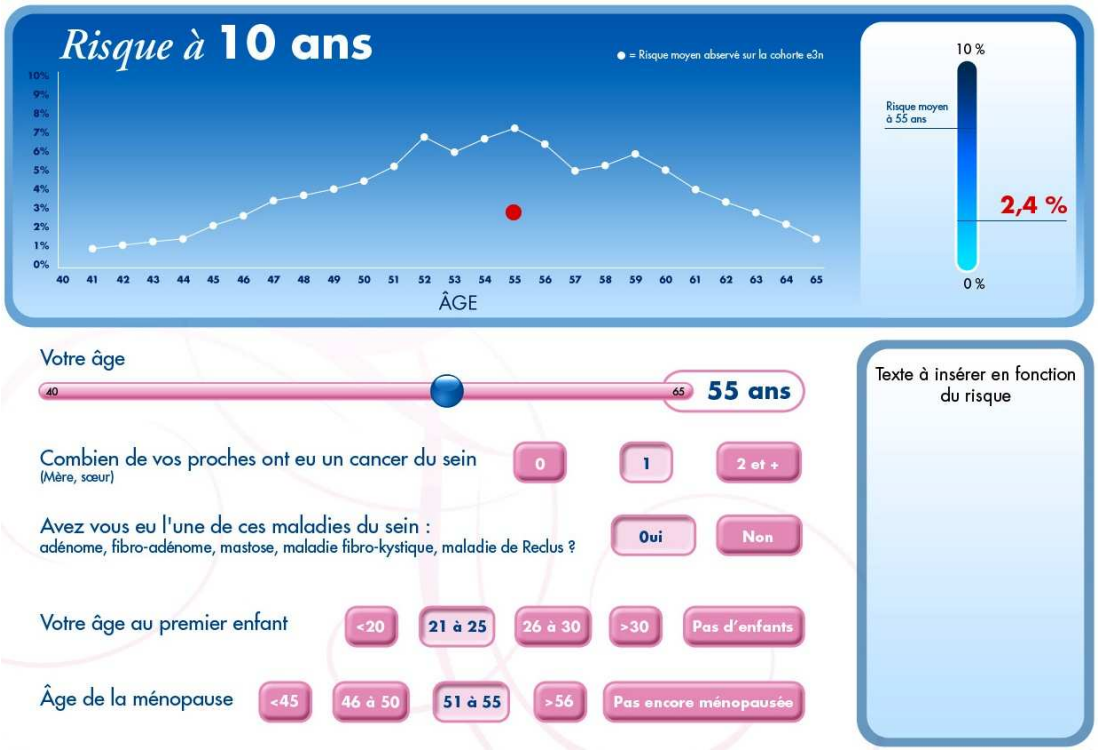

Figure 3.4 : Prototype d'application web pour afficher le risque de cancer du sein d'une femme à partir des données de l'étude E3N et des attributs retenus dans le score de risque.

sur le graphique sous la forme d'un rond (approximativement à 55 ans) et se déplace de manière instantanée quand les réponses aux questions du dessous sont modifiées par l'utilisateur.

Une fois validé par les responsables de la clinique du risque, ce prototype sera testé en conditions réelles au sein de l'IGR en 2013. Il sera intégré à un site web qui fournira des informations sur chaque facteur de risque pour le cancer du sein.

Dans ce chapitre, nous avons analysé le contexte métier pour fixer des objectifs en termes de proposition d'un processus de création de scores de risque et en termes d'utilisation d'une méthode de modélisation compréhensible. Nous avons montré les spécificités des caractéristiques des données utilisées pour la création d'un score de risque de santé en population générale en ce qui concerne le déséquilibre des données, la confidentialité et la possibilité d'action sur les attributs affichés dans le score de risque. Sur ces bases, nous avons proposé un processus de création de scores de risque basé sur le modèle de processus CRISP-DM dont nous montrons le déroulement complet sur le cas du cancer du sein dans un contexte de prévention au sein d'une clinique du risque.

Pour permettre de mener des expérimentations utilisant le processus proposé, nous avons conçu une architecture de système d'information détaillée dans le chapitre 4 et qui permet de produire des résultats détaillés dans le chapitre 5.

4 Architecture d'un système d'information supportant la production et l'utilisation de scores de risque

Avec l'objectif de mener à bien des expérimentations qui s'appuient sur une plate-forme réalisant les processus présentés dans le chapitre 3, nous proposons dans ce chapitre l'architecture d'un système d'information qui soutient les processus de production et d'utilisation de scores de risque. Après avoir détaillé l'intérêt de cette approche par rapport à nos travaux, nous concevons, sur la base de processus métiers que nous aurons modélisés, l'architecture fonctionnelle (i.e. fondée sur les fonctions attendues) du « système d'information du risque de maladie ». Enfin, nous mettons en œuvre des fonctions extraites de cette architecture sous la forme d'un système informatique (de l'architecture applicative au code) qui tient compte des contraintes métier.

4.1 Enjeux, définitions et objectifs

Dans cette partie, nous abordons les enjeux généraux à court et long termes de la conception d'un système d'information. Puis, nous détaillons nos objectifs en matière d'architecture dont celle d'un système d'information fonctionnellement exhaustif par rapport aux processus métiers. Le développement d'un système réalisant notre besoin métier est fondé sur l'extraction de certaines fonctions de ce système d'information. Ce besoin est l'évaluation du risque de cancer du sein au moyen des processus présentés au chapitre 3 grâce à un système d'information du risque de maladie.

4.1.1 Enjeux de la conception d'un système d'information

Dans le cadre de notre travail, l'enjeu principal de la conception d'un tel système d'information est l'automatisation de l'expérimentation ciblant les fonctionnalités (spécifiées dans les cas d'utilisation, partie 4.3.1, page 94) de production et d'utilisation de scores de risque dans le domaine de la santé. En effet, nous devons concevoir l'agencement des composants du système informatique supportant les processus métiers décrits dans le chapitre 3 afin de pouvoir expérimenter (chapitre 5). Le système d'information est caractérisé par une vue fonctionnelle qui est utilisable pour les développements de systèmes informatiques de ce système d'information.

La conception de la vue fonctionnelle du système d'information permet de mener

une réflexion sur les fonctions supportant les différentes tâches qui composent les processus métiers ciblés. Bien concevoir le système d'information et les systèmes qui le constituent, permet d'automatiser tout ou partie du déroulement des processus métiers. Par exemple, dans le cas de la production et de l'utilisation de scores de risque, un système d'information efficace et aligné sur les processus métiers [Simonin 11] identifiés comme utiles à l'expérimentation, permettra de dérouler plusieurs fois le processus métier identifié de création d'un score de risque. Les utilisateurs de ce processus auront des contraintes différentes :

- en termes d'attributs utilisables,
- en termes de données en fonction de leur d'origine et de leur composition,
- en termes de maladies différentes pour lesquelles il faudra prédire un niveau de risque.

La facilitation du déroulement d'un tel processus permettra de produire, rapidement et de façon répétée, des scores de risque adaptés à chaque contexte et chaque maladie tout en respectant les contraintes imposées.

4

4.1.2 Nos objectifs en matière d'architecture de système d'information et de système informatique

Notre objectif est de faciliter l'expérimentation de la création de scores de risque en l'automatisant au sein d'un système d'information du risque de maladie. Pour y parvenir, nous avons choisi de généraliser les processus mis en œuvre pour la création de scores de risque pour le cancer du sein afin de concevoir un système d'information en s'inspirant d'une méthode développée pour un système d'information de services de télécommunication [Simonin 12].

Au vu de la complexité algorithmique des fonctions à réaliser et du nombre de données à traiter du point de vue applicatif, l'utilisation de l'informatique permet un gain de temps en particulier dans l'assistance au choix des modèles de risque du fait de leur nombre.

Pour atteindre cet objectif grâce à cette méthode, nous avons utilisé différentes vues d'architecture pour décrire notre système d'information [Longépé 01].

- L'*architecture métier* couvre les processus mis en œuvre par les différents acteurs pour réaliser leurs métiers. Nous avons choisi de limiter les processus métiers à la conception de modèle et à l'obtention d'un niveau de risque à des fins d'automatisation de l'expérimentation. Elle est décrite partie 4.2, page 89.
- L'*architecture fonctionnelle* recense les fonctions qui devront être proposées par le système d'information. Elle est construite à partir des processus métiers et en partie déduite de l'analyse fonctionnelle de l'utilisation de ce système d'information. Elle est décrite en partie 4.3, page 94.
- L'*architecture applicative* est la mise en œuvre de l'architecture fonctionnelle d'un système informatique dans un environnement matériel et logiciel donné. Elle définit l'assemblage des composants applicatifs et leurs interfaces.
- L'*architecture technique* a pour objectif de détailler les solutions techniques

(nœud d'exécution* ou liens de communication* entre eux) qui supportent les processus métiers et qui permettent de déployer l'architecture applicative du système informatique.

Les architectures technique et applicative de notre système informatique sont décrites en partie 4.4, page 103.

4.1.3 Approche d'architecture fonctionnelle d'un système d'information

L'urbanisation des systèmes d'information [Sassoon 98] est une approche dont le but est de permettre l'évolution des systèmes d'information conformément à l'évolution de la stratégie d'entreprise. L'urbanisation des systèmes d'information permet de gérer le système d'information d'une entreprise pour assurer sa cohérence vis-à-vis des métiers de cette entreprise et des contraintes qui y sont associées. Les concepts de l'urbanisation des systèmes d'information sont calqués sur les concepts de l'urbanisation des habitats humains.

L'urbanisation doit permettre le partage de la connaissance du contenu d'un système d'information, mais ce partage doit également permettre le partage de concepts qui servent à décrire le contenu d'un système d'information ou d'un système informatique. Ces concepts sont décrits ici à l'aide de langages textuels (voir glossaire, page 153). Détaillons certains d'entre eux afin d'aider à la compréhension des termes et outils utilisés dans la suite du chapitre.

Dans notre approche, l'urbanisation du système d'information cible la conception de sa vue fonctionnelle. La conception est fondée sur l'alignement, ou recherche de cohérence, avec les processus* métiers autour du risque de maladie. Cet effort de traçabilité par rapport au métier, par alignement du système d'information fonctionnel sur les processus métier, doit participer à la notion de qualité de conception du système d'information [Comyn-Wattiau 10]. Cette conception permettra ensuite d'extraire certains îlots* de cette vue fonctionnelle ainsi que leurs relations (appelées voies dans le Plan Local d'Urbanisme des communes) afin de développer notre système.

En effet, pour faciliter l'évolution et la maintenance du système d'information, celui-ci est séparé en différentes parties appelées *îlots fonctionnels* regroupés en *quartiers fonctionnels*, eux-mêmes regroupés en *zones fonctionnelles*. Nous restreignons la conception du système d'information aux îlots fonctionnels, car seuls ceux-ci sont utiles au développement d'un système informatique. Les *données fonctionnelles** sont produites par un îlot fonctionnel et utilisées par un ou plusieurs îlots fonctionnels.

Le *diagramme de séquences fonctionnelles* permet d'illustrer des cas d'utilisation du système d'information. Ils intègrent les messages entre instances d'îlots fonctionnels et utilisent/produisent des données fonctionnelles en entrée/sortie. Ces diagrammes de séquences permettent d'établir les relations entre îlots fonctionnels et les relations entre données fonctionnelles.

L'utilisation du système d'information est décrite par des *cas d'utilisation**. Ceux-ci représentent les fonctionnalités offertes aux utilisateurs du système d'information [Booch 04]. La spécification des cas d'utilisation est le résultat d'une activité d'analyse des exigences fonctionnelles associées au système d'information. L'analyse des exigences fonctionnelles permet de structurer les besoins des utilisateurs en identifiant les utilisateurs du système et leurs interactions avec le système. Nous avons choisi d'aligner ces cas d'utilisation avec les tâches. Les *entités participantes* aux cas d'utilisation sont regroupées au sein d'un diagramme qui permet de visualiser les interactions entre ces entités participantes issues des scénarios illustrant les cas d'utilisation. Elles sont alignées avec les données métier et sont réalisées par les données fonctionnelles.

Au niveau du système informatique, nous concevons des *composants** *applicatifs* (vue organique) qui réalisent les îlots fonctionnels (extraits de l'architecture fonctionnelle du système d'information) et qui sont déployés sur des nœuds* d'exécution (architecture technique du système informatique).

Une relation entre deux composants applicatifs signifie ici l'utilisation d'une interface* d'un des deux composants par l'autre. Cette utilisation réalise une ou plusieurs relations entre îlots fonctionnels et est déployée sur un ou plusieurs liens* de communication.

4.2 Description des processus de modélisation et d'évaluation du risque d'une personne

Dans notre processus de conception du système d'information centré sur le risque, nous avons choisi de décrire deux processus métiers utiles à l'expérimentation. Nous avons donc généralisé les processus mis en œuvre, d'une part pour la conception d'un modèle de score de risque pour le cancer du sein, d'autre part pour l'obtention du niveau de risque d'une personne pour un modèle donné. Nous utilisons le langage xSPEM (« eXecutable Software Process Engineering Meta-Model ») [Bendraou 07] pour modéliser les processus métiers présentés.

4.2.1 Conception d'un modèle de risque

Afin de pouvoir mener des expérimentations pour tester notre processus de conception d'un modèle de risque pour le cancer du sein, nous avons identifié et spécifié du point de vue du métier, trois activités principales qui constituent le processus de conception d'un modèle de risque. Ce sont les activités de *conception d'une liste d'attributs* adaptée à la spécialité médicale et au contexte d'utilisation du score de risque, de *préparation des données* et de *choix du modèle du risque* qui inclut la génération des modèles et les mesures de leurs performances. Chacune de ces activités est ensuite subdivisée en une suite de tâches par exemple *sélectionner les attributs* ou *calculer les performances de risque*.

Les activités ont été choisies en fonction de leur importance dans le processus métier de la conception du modèle de risque. L'issue de chaque activité représente un jalon lors d'une instanciation du processus.

Chaque activité est représentée en une suite de tâches qui aboutit à la réalisation de l'activité.

- L'activité de *conception d'une liste d'attributs* par spécialité permet d'aboutir à une liste d'attributs qui soient effectivement disponibles dans une base de données épidémiologiques. Cette liste devant d'une part être filtrée par un épidémiologiste expert de la maladie étudiée et d'autre part validée par un analyste en charge du besoin du client, notamment en fonction de la spécialité médicale dans laquelle le score de risque sera amené à être utilisé.
- L'activité de *préparation des données* consiste à transformer les données brutes, qui correspondent à la liste validée des attributs produite précédemment, en une liste de données utilisables pour l'algorithme choisi dont dépendent certaines règles de gestion des données manquantes ou de discrétisation des données.
- Enfin, l'activité de *choix du modèle de risque* inclut la génération des différents modèles en fonction d'une configuration de l'algorithme choisi, en fonction :
  - d'une répartition des données pour permettre la validation croisée,
  - d'une liste de données précédemment construite, filtrée et validée.

  Pour chaque modèle généré, des indicateurs de performances sont automati-

quement générés : pour la mesure de discrimination, on génère la mesure d'aire sous la courbe ROC et le risque relatif observé tandis que pour la mesure de calibration, on génère le diagramme de fiabilité et le rapport du nombre estimé de cas sur le nombre observé de cas. En fonction du contexte, le choix du modèle à utiliser pourra soit être effectué par l'analyste et validé par l'épidémiologiste (voir partie 3.3.2.1, page 75 pour la description des parties prenantes), soit, dans la majorité des cas, être effectué par l'épidémiologiste expert du domaine et validé par l'analyste.

La description des activités métiers nous permet de les modéliser sous la forme de trois activités principales constituant le processus de conception d'un modèle de risque : elles sont présentées figure 4.1 d'après la description des processus faite en partie 3.3.2, page 75. Chacune des tâches fait intervenir une personne dont le rôle est indiqué sur la gauche du diagramme, par exemple *épidémiologiste* ou *analyste*. Chaque tâche peut utiliser des données métiers qualifiées d'entrantes et générer des données métiers qualifiées de sortantes qui peuvent éventuellement être utilisée comme données entrantes dans une tâche suivante. Une donnée métier est par exemple une *liste d'attributs* qui est une donnée sortante de la tâche *Filtrer les attributs* ou un *modèle de risque* qui est une donnée entrante de la tâche *Valider un modèle de risque*.

L'activité de *Conception d'une liste d'attributs* par spécialité est modélisée dans la figure 4.2, l'activité de *Préparation des données* en figure 4.3 et l'activité de *Choix du modèle de risque* en figure 4.4.

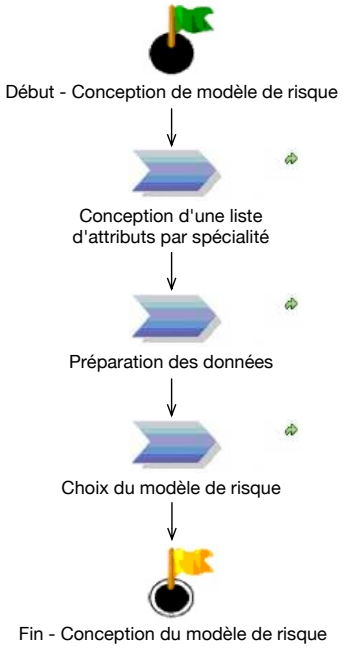

Figure 4.1 : Activités métiers du processus de conception du modèle de risque.

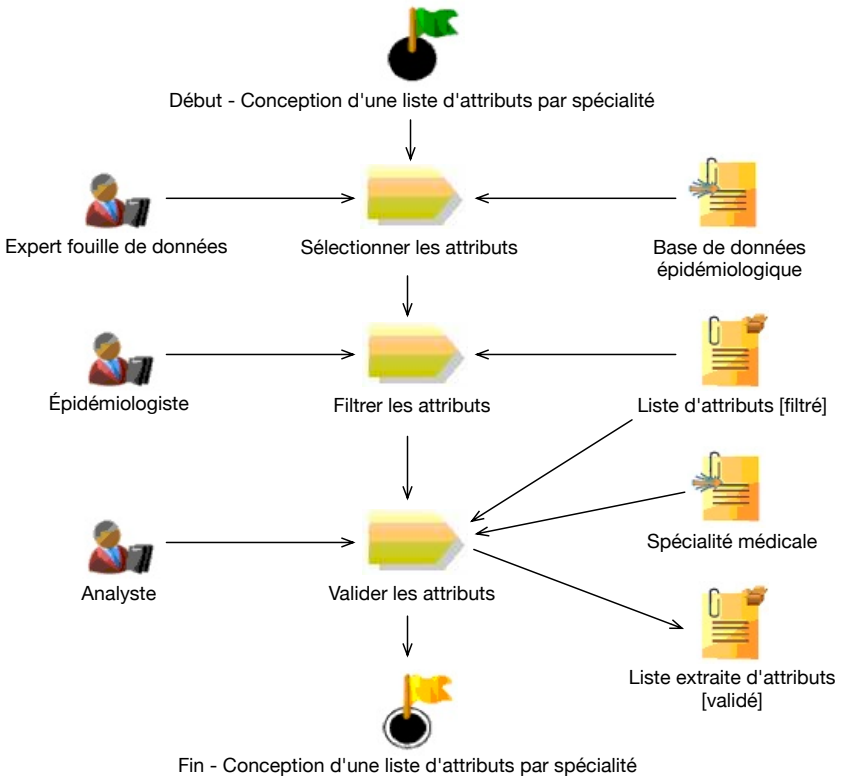

Figure 4.2 : Tâches métiers de l'activité de conception d'une liste d'attributs par spécialité.

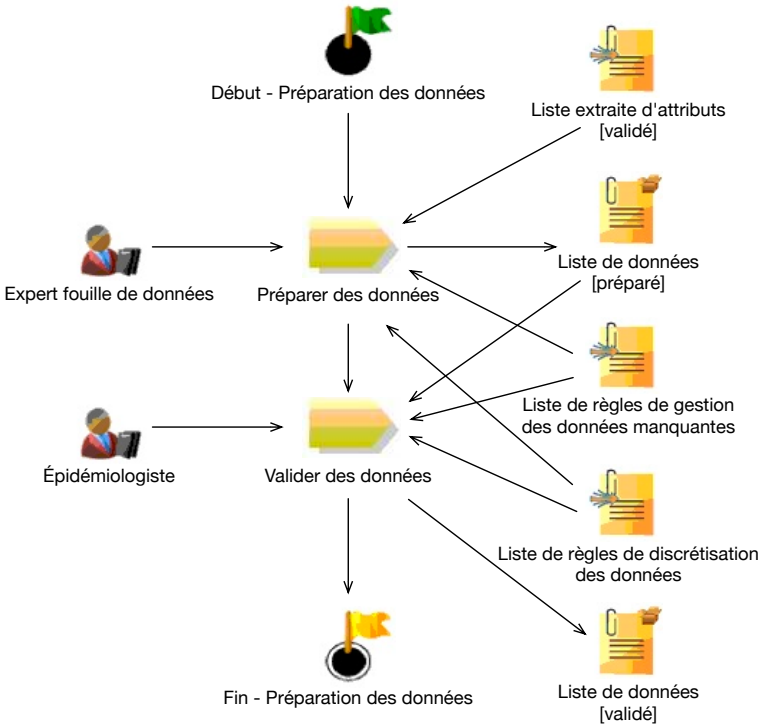

Figure 4.3 : Tâches métiers de l'activité de préparation des données.

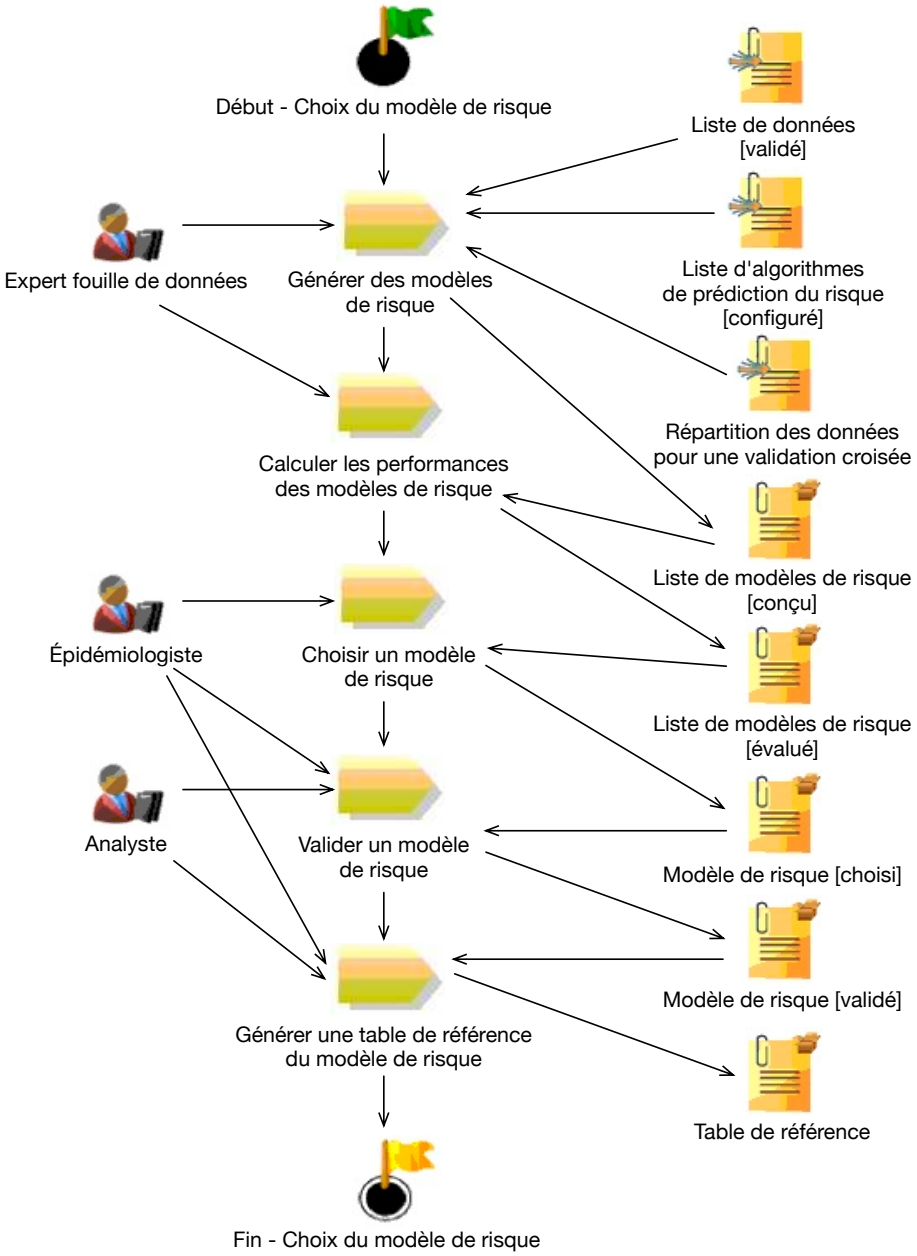

Figure 4.4 : Tâches métiers de l'activité de choix du modèle de risque.

4.2.2 Obtention niveau de risque d'une personne

Le processus métier d'obtention du niveau de risque est constitué d'une seule activité dont l'objectif est d'*obtenir le niveau de risque d'une personne*. La première tâche de l'activité est la *Saisie du profil d'une personne*. Ce profil sera évalué en utilisant les scores précalculés au cours du processus de *Conception du modèle de risque* et stocké sous la forme d'une donnée métier appelée *table de référence* qui sera utilisée dans le processus d'*obtention du risque d'une personne*.

L'utilisation d'une table de référence permet de répondre à une contrainte essentielle de la construction d'un score de risque à partir de données de type médicale qui ne doivent pas être rendues publiques. La table de référence est conçue à partir du modèle choisi, indépendamment de la méthode choisie pour modéliser le risque et permet d'associer un niveau de risque à un profil de personne sans stockage d'informations personnelles. De plus, dans cette activité, nous avons fait le choix de ne pas différencier le processus en fonction du rôle d'utilisateur du système d'évaluation du niveau de risque, qu'il soit médecin ou patient. En effet, le processus est identique pour le médecin et pour le patient.

De la description de ce processus métier, nous déduisons le modèle d'activité correspondant. Cette activité est détaillée dans la figure 4.5.

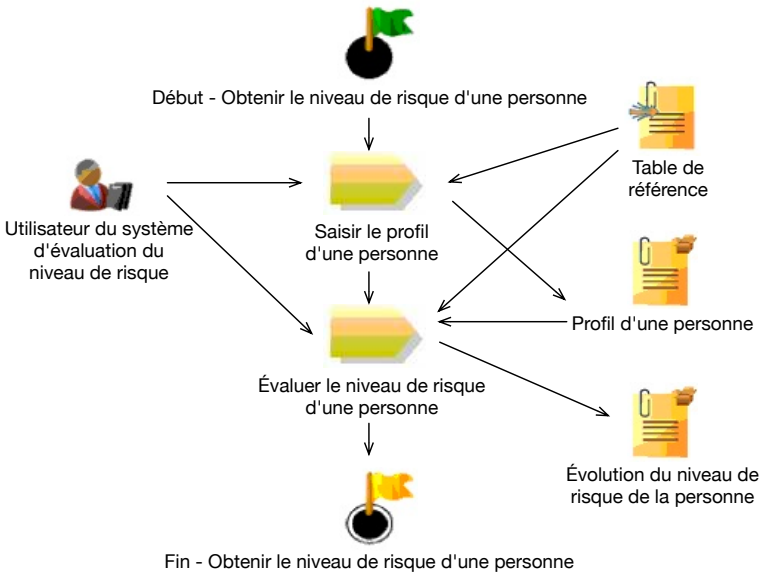

Figure 4.5 : Tâches métiers de l'activité d'obtention du niveau de risque d'une personne.

Nous avons choisi de fonder l'architecture fonctionnelle du système d'information sur son alignement avec les deux processus métiers détaillés au niveau des tâches et des rôles qui en sont responsables.

4.3 Conception fonctionnelle du système d'information du risque de maladie

La conception du système d'information du risque de maladie consiste en la conception de sa vue fonctionnelle indépendamment de sa couverture informatique*. Dans cette partie, nous présentons la conception fonctionnelle du système d'information déduite d'une spécification de ses cas d'utilisation. Afin que la conception soit cohérente avec les processus métiers spécifiés, nous avons choisi d'aligner ces cas d'utilisation avec les tâches composant les processus métiers présentés dans la partie précédente (partie 4.2).

4.3.1 Spécification fonctionnelle du système d'information du risque de maladie fondée sur les processus

La spécification fonctionnelle du système d'information est présentée en deux étapes : les cas d'utilisation sont présentés puis les entités participantes.

4.3.1.1 Cas d'utilisation

Les cas d'utilisation ont été conçus par alignement avec les tâches métiers décrites précédemment : à chaque tâche, nous avons fait correspondre un cas d'utilisation dont la description se trouve en annexe A.1 du manuscrit, page 161. Le tableau 4.1 présente ces correspondances. Chaque cas d'utilisation a été nommé en fonction du nom de l'activité sur laquelle il est aligné, en ajoutant le préfixe « Cu » qui signifie « cas d'utilisation ».

Tableau 4.1 : Alignement des tâches métiers avec les cas d'utilisation du système d'information du risque de maladie.

Tâche	Cas d'utilisation
Sélectionner les attributs	CuSélectionAttributs
Filtrer les attributs	CuFiltrageAttributs
Valider les attributs	CuValidationAttributs
Préparer les données	CuPréparationDonnées
Valider les données	CuValidationPréparation
Générer des modèles	CuGénérationModèles
Calculer les performances des modèles	CuCalculPerformances
Choisir un modèle	CuChoixModèle
Valider un modèle	CuValidationChoixModèle
Générer une table de référence	CuGénérerTableRéférence
Saisir le profil d'une personne	CuSaisirProfil
Évaluer le niveau de risque	CuÉvaluerRisque

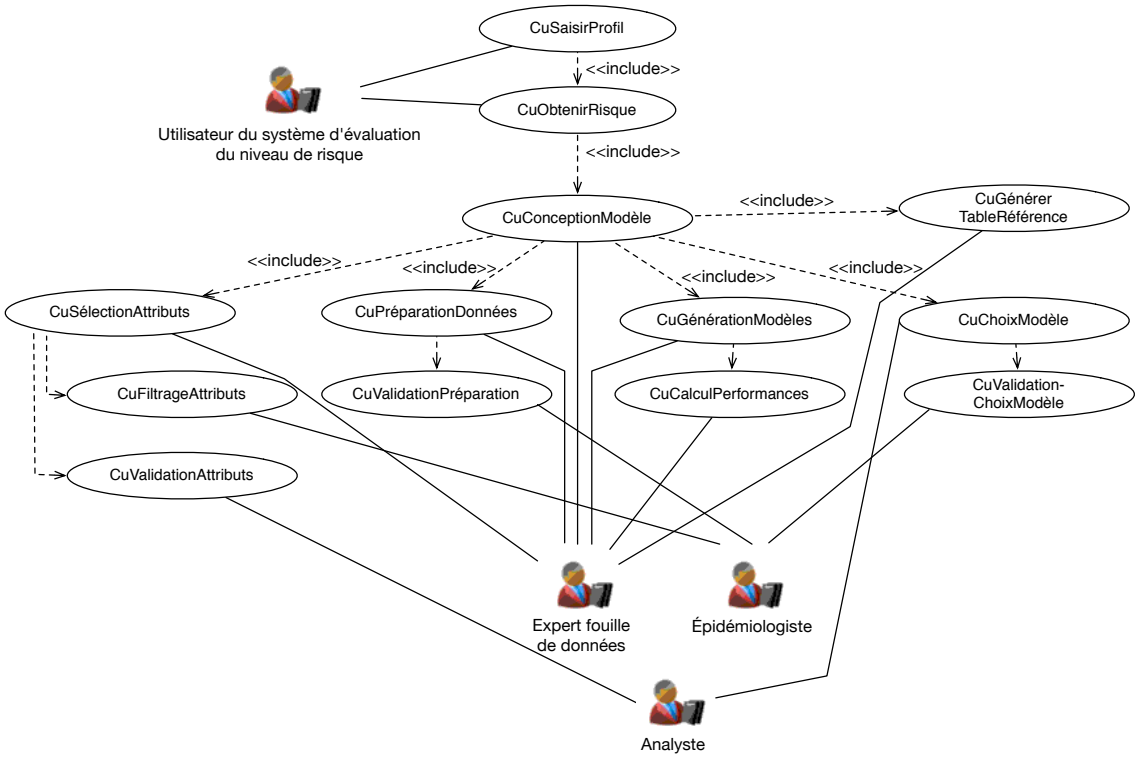

Figure 4.6 : Diagramme des cas d'utilisation du système d'information supportant la production et l'utilisation de scores de risque.

Le diagramme reprenant les cas d'utilisation est présenté en figure 4.6. Il permet de visualiser les relations qui existent entre eux, (de type « include* »). Les cas d'utilisation mis en jeux lors de la conception d'un modèle de risque se lisent de gauche à droite pour respecter l'ordre des activités montrées dans le processus *Conception d'un modèle de risque*, partie 4.2, page 89. Ce diagramme des cas d'utilisation est aligné avec les diagrammes représentant les tâches de chaque activité (figures 4.2 à 4.5) : le rôle responsable de la tâche est en interaction avec le cas d'utilisation et les liens entre cas d'utilisation de type « include » sont déduits des séquences de tâches. Par exemple, la succession des tâches *Préparer les données* précédant *Valider les données* est alignée avec le lien d'inclusion de *CuPréparationDonnées* vers *CuPréparationValidation*.

Les rôles responsables de ces tâches sont alignés avec la description de l'activité : successivement *expert en fouille de données*, *épidémiologiste* et *analyste*. Par exemple, le rôle d'*expert en fouille de données* est responsable de la tâche interagit avec le cas d'utilisation *CuPréparationDonnées*.

Tableau 4.2 : Entités participantes aux cas d'utilisation.

Nom de l'entité	Définition de l'entité
Base de données	Contient l'ensemble des attributs utilisables pour construire le score de risque.
Maladie	Maladie pour laquelle on cherche à construire un score de risque.
Spécialité médicale	Contient le nom de la spécialité médicale pour laquelle le score est construit
Attribut	Représente un facteur de risque pour la maladie dont on cherche à évaluer le risque.
Donnée	Contient la valeur correspondant à un attribut.
Règle	Contient les informations sur la manière de nettoyer les données
Configuration	Représente un paramétrage de l'algorithme utilisé.
Combinaison	Constituée d'attributs, elle constitue un élément de la configuration.
Répartition	Contient la répartition des exemples dans les jeux d'apprentissage ou de test.
Algorithme	Contient une liste des paramètres propre à l'algorithme utilisé pour construire le modèle.
Modèle	Construit à partir d'une configuration, il représente le moyen d'évaluer un risque de maladie. Il est caractérisé par des mesures de performance.
Sélectionneur	Représente la personne qui sera chargée de choisir un modèle parmi d'autres.
Validateur	Représente la personne qui sera chargée de valider le choix du modèle.
Table de référence	Contient l'association entre les profils d'individus et les niveaux de risque
Niveau de risque	Représente le niveau de risque prédit pour une maladie pour un profil à partir d'un modèle.
Utilisateur	Représente l'utilisateur du système qui apporte le profil dont il faut évaluer le risque

4.3.1.2 Entités participantes

Les entités participantes sont extraites de l'ensemble des cas d'utilisation. Ces entités participantes sont définies dans le tableau 4.2.

Le niveau de détail dans la spécification d'une entité participante est justifié par son importance au niveau métier, importance déduite de la description des processus métiers faite au chapitre 3. Par exemple, la conception de modèle étant un processus métier dont le résultat dépend étroitement de la configuration particulière d'un algorithme utilisé, nous avons choisi de détailler l'entité *Configuration* en trois entités plus précises qui peuvent séparément influer les modèles générés, et donc le modèle conçu. Cette entité est donc détaillée en une entité *Combinaison* qui contient les informations nécessaires à la construction des combinaisons d'attributs dont il faut tester la capacité de prédiction du niveau de risque et dont un paramètre important est le nombre d'attributs, en une entité *Répartition* qui permet la prise en compte de la validation croisée afin d'estimer la validité statistique des résultats de prédiction de risque et en une entité *Algorithme* qui regroupe l'algorithme utilisé et les paramètres choisis pour le faire fonctionner.

Pour visualiser les relations entre les entités participantes, nous proposons le diagramme de la figure 4.7 qui reprend les entités participantes du tableau 4.2 en utilisant le formalisme du langage de modélisation UML [Booch 04]. De la même façon que pour les entités participantes, la relation entre entités est justifiée par leur

signification métier. Par exemple, la relation entre le *Modèle* et la *Table de référence* est justifiée par le fait que la *Table de référence* est générée par le *Modèle*, cette dernière contenant les scores précalculés par le modèle.

Ce diagramme comporte au centre l'entité *Modèle* ce qui est cohérent avec la conception fonctionnelle d'un système d'information ayant pour but de faciliter l'expérimentation dans la conception de modèles de risque. Il est à noter que l'entité *Attribut* est à la fois attachée aux bases de données qui sont utilisées pour construire les modèles et à la fois attachée aux utilisateurs qui fournissent au système, le profil d'une personne constitué de valeurs attachées à des attributs. Cette double appartenance aura des conséquences sur l'architecture fonctionnelle.

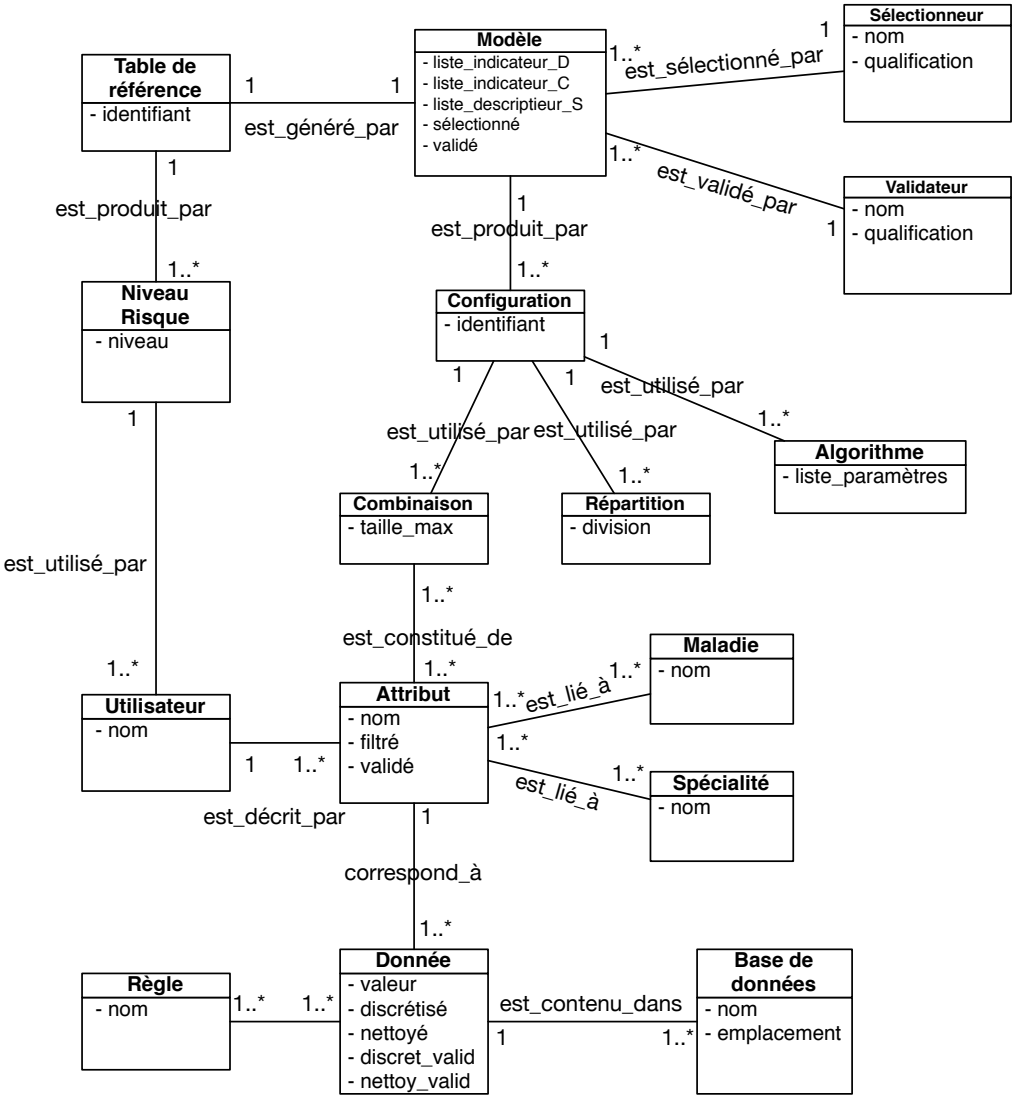

Figure 4.7 : Diagramme des entités participantes aux cas d'utilisation du système d'information supportant la production et l'utilisation de scores de risque.

4.3.2 Architecture fonctionnelle du système d'information du risque de maladie

L'architecture fonctionnelle du système d'information est déduite de son analyse fonctionnelle (voir la définition des termes page 86). Cette architecture fonctionnelle cible la conception des îlots fonctionnels, des données fonctionnelles, qu'ils produisent et utilisent, et des relations entre îlots fonctionnels et entre données fonctionnelles.

4.3.2.1 Conception des îlots fonctionnels de gestion du risque de maladie et des données fonctionnelles produites

La conception d'îlots fonctionnels est une activité d'urbanisme des systèmes d'information. Elle consiste en un découpage du système d'information en îlots de périmètres pertinents. Le périmètre d'un îlot correspond au nombre de fonctions regroupées. L'objectif est de faciliter l'intégration fonctionnelle de systèmes informatiques et de faciliter des évolutions fonctionnelles ciblées du système d'information.

Les îlots fonctionnels sont conçus à partir des entités participantes aux cas d'utilisation et de leurs relations. Les choix de conception (par exemple le nombre d'îlots fonctionnels et leur périmètre) sont notamment guidés par la connaissance du métier et la cible poursuivie, dans notre cas, la gestion du risque de maladie.

Le tableau 4.3 montre la manière dont nous avons travaillé pour concevoir les îlots fonctionnels : chaque attribut, des entités participantes aux cas d'utilisation du système, doit être réalisé par une donnée fonctionnelle produite par un îlot fonctionnel. Ce tableau présente ces îlots fonctionnels, les attributs associés aux entités participantes et les données fonctionnelles extraites de ces entités et produites par les îlots de gestion du risque.

Le fait de guider les choix de conception par la connaissance du métier apparaît notamment sur le périmètre que nous avons défini pour les îlots de gestion. Par exemple, en fonction de notre connaissance de la répartition des compétences, nous avons séparé les îlots relevant des compétences des épidémiologistes et les îlots qui relèvent de la compétence de la fouille de données. Cela se traduit par un îlot de gestion des données qui regroupe les entités base de données, maladie, attribut, spécialité et données. Cet îlot de gestion des données est distinct de l'îlot de configuration des algorithmes, car les compétences sont différentes entre les épidémiologistes et les spécialistes d'algorithmes de fouille de données. De même, c'est la connaissance du métier qui nous conduit à séparer la gestion des calculs et des choix qui sont faits alors même qu'ils sont regroupés dans un processus. Les compétences demandées aux acteurs du processus sont en effet très différentes puisqu'il s'agit d'une part d'automatiser la production de modèles de risque et d'autre part d'enregistrer des choix effectués par les experts du domaine comme les épidémiologistes.

Tableau 4.3 : Îlots fonctionnels et attributs associés aux entités participantes.

Entité participante	Attributs	Îlot fonctionnel	Données fonctionnelles
Base de données	nom, emplacement	IF Gestion données	DF Base de données
Maladie	nom	IF Gestion données	DF Maladie
Spécialité médicale	nom	IF Gestion données	DF Spécialité
Attribut	nom, filtré, validé	IF Gestion données	DF Attribut
Données	valeur, discrétisé, nettoyé, discret_valid, nettoy_valid	IF Gestion données	DF Données
Règle	nom	IF Gestion règle	DF Règle
Configuration	identifiant	IF Gestion configuration	DF Configuration
Combinaison	taille_max	IF Gestion configuration	DF Combinaison
Répartition	division	IF Gestion configuration	DF Répartition
Algorithme	liste_paramètres	IF Gestion configuration	DF Algorithme
Modèle	liste_indicateurs_discrimination, liste_indicateurs_calibration, liste_descripteurs_statistiques, sélectionné, validé	IF Gestion calcul	DF Modèle
Table de référence	identifiant	IF Gestion calcul	DF Table de référence
Niveau de risque	niveau	IF Gestion calcul	DF Niveau de risque
Sélectionneur	nom, qualification	IF Gestion choix	DF Sélectionneur
Validateur	nom, qualification	IF Gestion choix	DF Valideur
Utilisateur	nom	IF Gestion utilisateur	DF Utilisateur

4.3.2.2 Relations entre îlots et relations entre données utiles à la gestion du risque de maladie

Afin de définir les relations entre îlots fonctionnels d'une part et entre données fonctionnelles d'autre part, nous concevons les diagrammes de séquences qui représentent les scénarios qui illustrent les cas d'utilisation. La figure 4.8 montre un diagramme de séquence important : il s'agit du diagramme de séquence correspondant à la conception d'un modèle de risque qui résume le fonctionnement de la création d'un modèle de risque. Ce diagramme correspond au scénario nominal (hors exception) du cas d'utilisation *CuConceptionModèle* dont il détaille les messages échangés et leur séquence temporelle. Les opérations, comme *ObtenirConfiguration*, sont des parcelles* de l'îlot fonctionnel *IF Gestion configuration*.

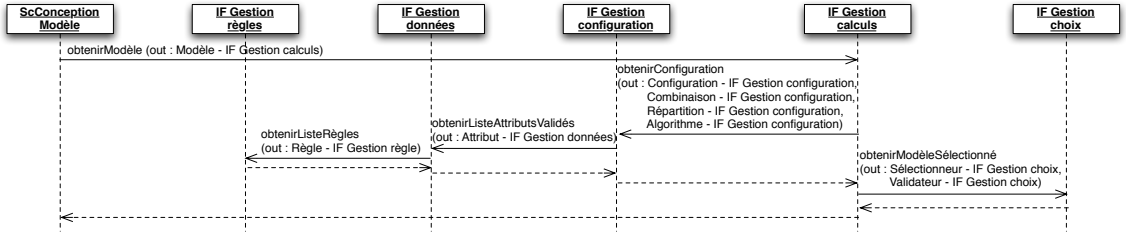

Figure 4.8 : Diagramme de séquence du scénario nominal du cas d'utilisation *CuConceptionModèle*.

Les diagrammes de séquences correspondants aux scénarios illustrant les autres cas d'utilisation sont référencés dans le tableau 4.4. Ce tableau montre la correspondance entre chaque cas d'utilisation et chaque scénario ainsi que la référence du diagramme de séquence correspondant au scénario.

À partir des messages échangés entre instances d'îlots fonctionnels dans les diagrammes de séquences, nous déduisons les relations entre îlots fonctionnels. Un message de requête (flèche pleine dans le diagramme 4.8) envoyé d'une instance d'un îlot fonctionnel vers une instance d'un autre îlot fonctionnel induit une relation de dépendance du premier îlot fonctionnel vers le second. Ces relations sont schématisées sous la forme d'un diagramme de composants [Booch 04] présenté en figure 4.9.

Tableau 4.4 : Correspondance entre cas d'utilisation, scénarios et position du diagramme de séquence correspondant en annexe.

Cas d'utilisation	Scénario	Diagramme de séquence
CuSélectionAttributs	ScSélectionAttributs	Figure A.1, page 168
CuFiltrageAttributs	ScFiltrageAttributs	Figure A.2, page 168
CuValidationAttributs	ScValidationAttributs	Figure A.3, page 169
CuPréparationDonnées	ScPréparationDonnées	Figure A.4, page 169
CuValidationPréparation	ScValidationPréparation	Figure A.5, page 169
CuCalculPerformances	ScCalculPerformances	Figure A.6, page 170
CuChoixModèle	ScChoixModèle	Figure A.7, page 170
CuValidationChoixModèle	ScValidationChoixModèle	Figure A.8, page 171
CuGénérerTableRéférence	ScGénérerTableRéférence	Figure A.9, page 171
CuSaisirProfil	ScSaisirProfil	Figure A.10, page 172
CuObtenirRisque	ScObtenirRisque	Figure A.11, page 172

Ensuite, à partir des diagrammes de séquence et des données fonctionnelles, nous déduisons les relations entre ces données. Chaque dépendance entre données fonctionnelles doit en effet être cohérente (existence et orientation) avec la dépendance entre les îlots fonctionnels qui produisent ces données. Par exemple la dépendance de la donnée fonctionnelle *Combinaison - IF Gestion configuration* vers la donnée

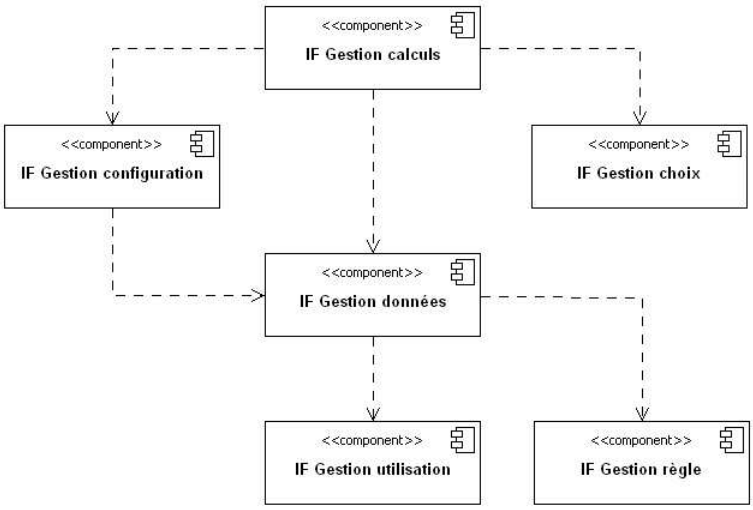

Figure 4.9 : Diagramme des îlots fonctionnels du système d'information supportant la production et l'utilisation de scores de risque.

fonctionnelle *Attribut - IF Gestion données* est cohérente avec la dépendance de l'îlot fonctionnel *IF Gestion configuration* vers l'îlot fonctionnel *IF Gestion données* visible sur la figure 4.9. La dépendance de la donnée fonctionnelle *Attribut - IF Gestion données* vers la donnée fonctionnelle *Règles - IF Gestion règle*, visible sur la figure 4.10, est déduite du diagramme de séquence de la figure 4.8.

Ensuite, à partir des diagrammes de séquences et des données fonctionnelles, nous déduisons les relations entre données fonctionnelles. Ces relations sont schématisées sous la forme d'un diagramme de classe [Booch 04] présenté en figure 4.10. Le nom de chaque classe UML correspondant à une donnée est constitué du nom de l'entité dont elle est produite, un tiret puis le nom de l'îlot fonctionnel qui la produit.

Il est intéressant de noter que la dualité, évoquée page 97, de la donnée fonctionnelle *Donnée - IF Gestion données* qui regroupe à l'information utilisée pour construire les modèle et l'information qui permet de décrire le profil à évaluer une fois le modèle de risque choisir, persiste logiquement dans ce diagramme. Cette donnée fonctionnelle est reliée à la fois à la donnée fonctionnelle liée aux calculs pour construire le modèle de risque et aussi à la donnée fonctionnelle liée à l'utilisation du score de risque dans lequel elle permet de décrire le profil d'un utilisateur.

L'étape suivante consiste en un développement de systèmes réalisant les îlots fonctionnels dédiés à l'évaluation du risque. Les composants applicatifs de ce système réalisent des îlots fonctionnels du système d'information du risque de maladie et leurs relations.

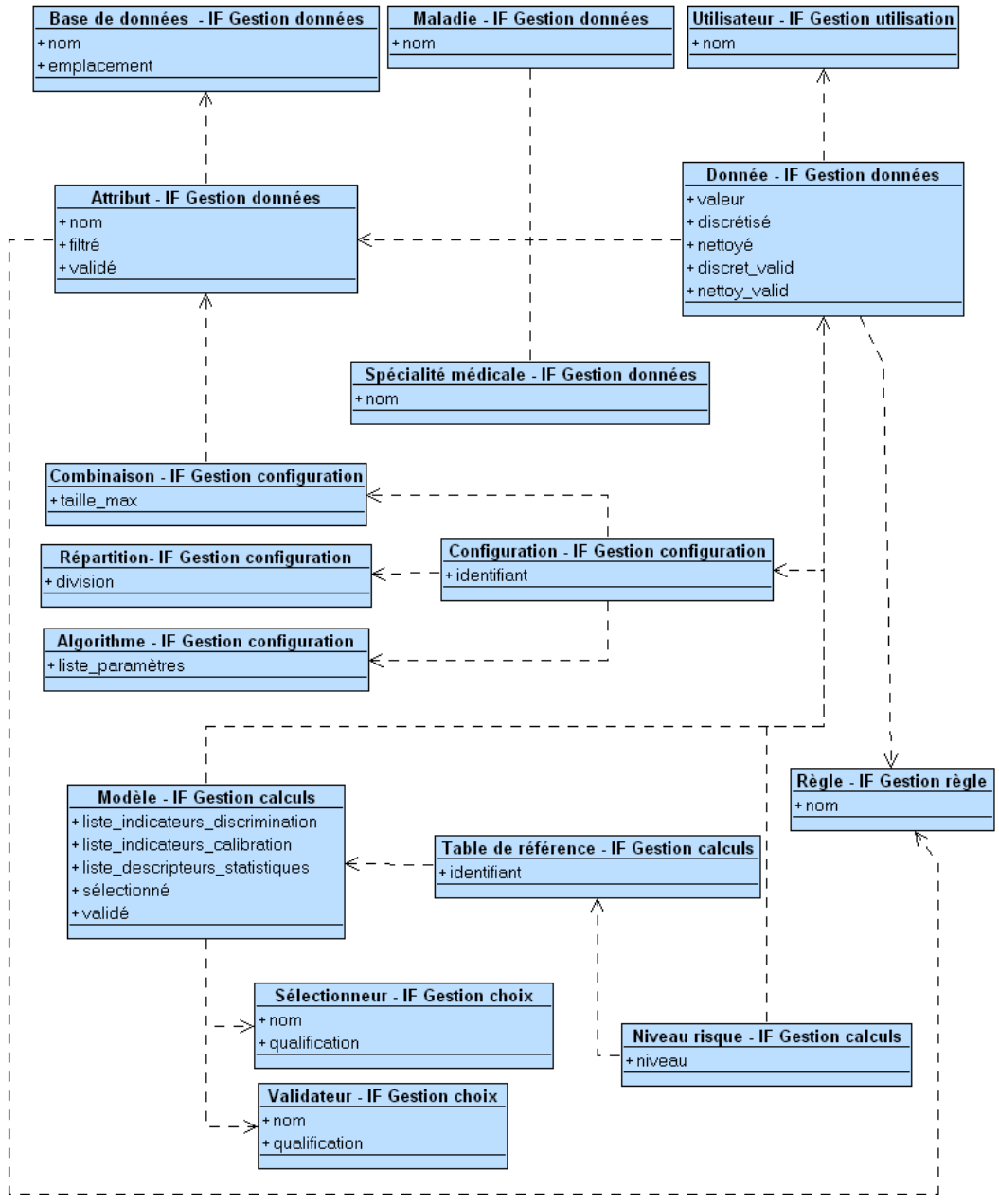

Figure 4.10 : Diagramme des données fonctionnelles du système d'information supportant la production et l'utilisation de scores de risque.

4.4 Système informatique de mise à disposition du niveau de risque

Le développement du système informatique de mise à disposition du niveau de risque est fondé sur la réalisation applicative d'un extrait de la vue fonctionnelle du système d'information (voir partie 4.3.2, page 98). Notre objectif est de concevoir des composants applicatifs permettant de réaliser des expérimentations de fouille de données afin de construire un score de risque de maladie. Dans ce but, nous présentons dans un premier temps l'architecture des éléments applicatifs, puis nous soulignons la stratégie, fondée sur les connaissances métiers, mise en œuvre pour la concevoir. Enfin, nous présentons la mise en œuvre du composant applicatif de gestion des calculs due à son importance métier.

4.4.1 Architecture applicative du système informatique de mise à disposition du niveau de risque

S'il arrive que ce soit des contraintes d'ordre technique qui imposent certains choix de conception d'architecture applicative tel que dans le cycle de développement décrit dans [Simonin 12], dans le cas de notre système, certains aspects de l'architecture applicative sont contraints par le métier. Par exemple, le fait d'utiliser des données médicales, qui sont entreposées dans le système d'information d'un hôpital, impose de fortes restrictions quant aux liens de communications qui peuvent être créés pour relier les nœuds d'exécution conçus en architecture technique. Notamment, la politique de sécurité informatique de l'hôpital interdit formellement d'automatiser un transfert de données d'un nœud d'exécution situé en son sein vers un nœud d'exécution extérieur.

Ces contraintes doivent donc être prises en compte dans l'architecture applicative du système informatique. Cette architecture est présentée sous la forme de composants applicatifs réalisant des îlots fonctionnels et déployés sur des nœuds d'exécution.

La figure 4.11, présente les composants applicatifs déployés sur les nœuds d'exécution à disposition. Chaque composant applicatif est nommé selon l'îlot fonctionnel qu'il réalise auquel on accole le nom des données fonctionnelles correspondantes. En effet, nous avons choisi de définir chaque composant applicatif par rapport à la réalisation de la gestion des données fonctionnelles utiles à notre système informatique. Les flèches pointillées représentent des utilisations d'interfaces : le composant à l'origine de la dépendance UML utilise une interface du composant à l'extrémité de la dépendance. Par exemple, le composant applicatif *CA Gestion Calcul - Niveau Risque* utilise une interface du composant applicatif *CA Gestion calculs - Modèle + Table de Référence*. Les traits pleins représentent les liens de communication entre les nœuds d'exécution. Par exemple, le lien de communication entre le *Serveur Web* et le *Serveur de calcul* où est déployée l'utilisation de l'interface applicative précédente.

Ces nœuds d'exécution sont décrits ci-dessous.

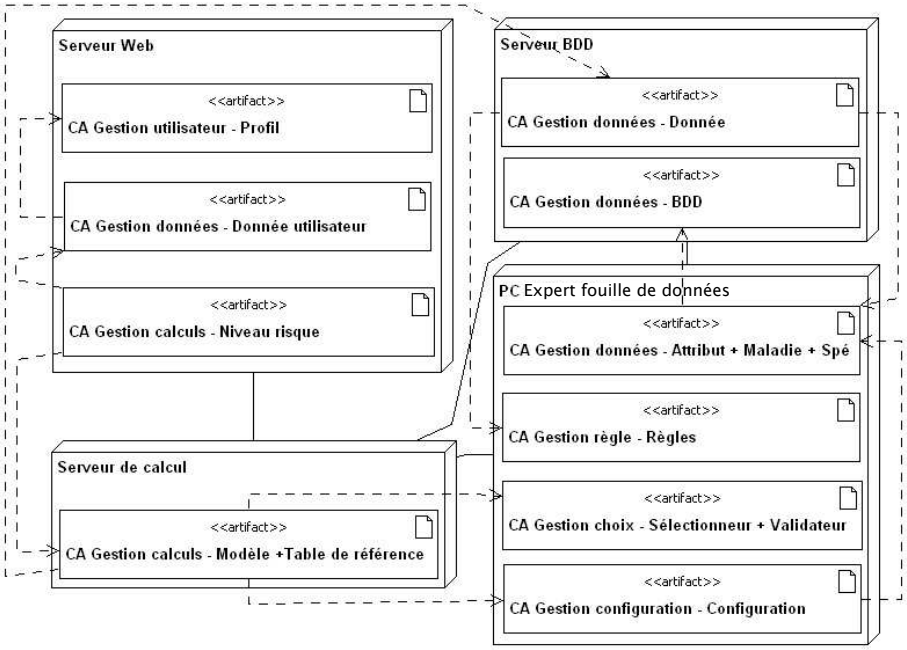

Figure 4.11 : Modèle de déploiement applicatif du système informatique de mise à disposition du niveau de risque.

- Le *Serveur de Base De Données* (serveur BDD sur la figure 4.11) permet d'exécuter les composants de gestion de données relatifs aux bases de données et aux données métiers elles-mêmes.
- Le *PC de l'expert en fouille de données* est chargé de faciliter la coordination entre les différents nœuds d'exécution. Il permet notamment de réaliser manuellement les interfaces qui ne peuvent être réalisées automatiquement entre le serveur de calcul et le serveur de bases de données. Il permet, grâce aux composants applicatifs de gestion de données et de règles, de gérer l'étape de préparation des données, décrite dans le processus proposé au chapitre 3. Grâce aux composants applicatifs de gestion de configuration et des choix, il permet d'enregistrer les différents choix effectués :
  - lors de la phase de sélection des attributs utilisés pour modéliser le risque,
  - suite à la préparation des données,
  - suite à la génération de modèles effectuée par le serveur de calcul.
- Le *Serveur Web* gère l'interface avec l'utilisateur avec comme objectif l'obtention d'un profil. Il est en charge de la gestion des données pour la partie utilisateur de celles-ci et de gestion des calculs pour ce qui est du niveau de risque (correspondant au profil utilisateur entré comme donnée utilisateur) qui est extrait de la table de référence. Il correspond à la partie déploiement et utilisation du processus décrit au chapitre 3. Il possède un lien de communication avec le serveur de calcul afin d'autoriser l'échange de données telles que

la table de référence.

– Le *Serveur de calcul* est en charge du composant applicatif de gestion des calculs qui gère la génération des modèles et d'une table de référence. Le composant *CA Gestion calculs - Modèle + Table de référence* a une importance métier particulière, car c'est le composant chargé d'automatiser la génération des différentes modélisations possibles du risque de maladie qui découlent des choix d'attributs et de la préparation des données (voir la modélisation du processus de Conception de modèle, page 89).

4.4.2 Mise en œuvre du système informatique de mise à disposition de niveau de risque

Du fait de l'importance métier du composant *CA Gestion calculs - Modèle + Table de référence* (le chapitre 5 est dédié à l'analyse des modèles générés par ce composant), nous détaillons la description de la mise en œuvre de ce composant. L'objectif est de souligner l'importance du métier lors de cette phase de mise en œuvre du système informatique.

L'algorithme 1 présente une vision haut niveau de la partie du composant applicatif qui conduit à la production de la donnée *Liste de modèles* lors des activités *Générer des modèles de risque* et *Calculer les performances des modèles de risque*.

La partie conduisant à la production de la donnée *Table de référence* par le même composant est réalisée par un algorithme d'extraction de valeurs à partir d'un modèle qui aura été choisi conformément au processus modélisé dans ce chapitre page 92.

Les données fournies en entrée sont produites par les autres composants applicatifs décrits figure 4.11.

À partir de ces données, l'algorithme permet au composant applicatif de générer de manière automatique plusieurs dizaines à plusieurs centaines de modélisations du risque de maladie en fonction de la puissance de calcul disponible, et cela pour différents jeux de données, différents algorithmes avec leurs configurations, différentes combinaisons d'attributs (extraits des données reçues en entrée) et différentes répartitions de validation croisée.

En sortie, l'algorithme retourne une liste de modèles qui sont dans l'état *[évalué]*. Ces modèles associent les valeurs de scores qui constituent le modèle lui-même aux statistiques descriptives qui sont générées sur ces valeurs de scores et aux performances qui ont été évaluées.

Algorithme 1: Fonctionnement du composant *CA Gestion calculs - Modèle + Table de référence.*

Données : Liste de données [validé] (ListeDonnées),
Liste d'algorithmes (ListeAlgorithme),
Liste de configurations d'algorithme (ListeConfigAlgo),
Liste de répartitions validation croisée (ListeRépartition)
Résultat : Liste de modèles [évalué]

ListeModèles ⟵ ∅
ListeCombinaison ⟵ Générer les combinaisons d'attributs (ListeDonnées)

pour *chaque algorithme* algo ∈ ListeAlgorithme **faire**
  pour *chaque configuration* config ∈ ListeConfigAlgo **faire**
    pour *chaque combinaison* combi ∈ ListeCombinaison **faire**
      pour *chaque répartition* rep ∈ ListeRépartition **faire**
        ModèlePartiel ⟵ Générer un modèle(ListeDonnées, algo, config, combi, rep)
        StatistiquesDescriptivesPartielles ⟵ Générer des statistiques (ModèlePartiel)
        ÉvaluationPerformancesPartielles ⟵ Évaluer les performances (ModèlePartiel)
      fin
      Modèle ⟵ Combiner les modèles partiels (ModèlePartiel, StatistiquesDescriptivesPartielles, ÉvaluationPerformancesPartielles)
      ListeModèles ⟵ ListeModèles + Modèle
    fin
  fin
fin
Retourner *ListeModèles*

Le tableau 4.5 présente les correspondances entre les opérations des composants applicatifs, les classes contenant le code écrit pour le développement du composant applicatif et les étapes correspondantes de l'algorithme. Il est intéressant de noter que la fonction de ce composant est de mesurer de la capacité de prédiction de différents algorithmes avec différents paramétrages (au niveau de la configuration, des combinaisons, de la répartition des jeux d'apprentissage et de test), sans priorité d'un test sur un autre. Ces tests sont donc aisément parallélisables. Nous avons appelé ces différents niveaux de parallélisation des unités.

Les classes dont le préfixe du nom est *Unite* sont chargées de distribuer la charge de calcul dans des fils* d'exécution. Grâce au langage Java que nous avons utilisé (version 1.6), il est possible d'adapter automatiquement le nombre de fils exécutés simultanément aux capacités offertes par le nœud d'exécution sur lequel le composant est utilisé. Pour cela, toutes les classes sont instanciées et envoyées à un gestionnaire de fils d'exécution qui se charge d'exécuter les instances en respectant une limite d'instances exécutées simultanément à ne pas dépasser et en regroupant les résultats dans une instance de classe qui n'est accessible que lorsque celles-ci ont toutes ont été exécutées.

Tableau 4.5 : Correspondance entre les opérations des composants applicatifs, les classes et les étapes de l'algorithme 1.

Opération	Classe	Étape
Génération des combinaisons d'attributs qui seront testées pour prédire un risque	GenererCombinaisons	Générer les combinaisons d'attributs
Parcours de la liste des algorithmes à tester	UniteParcoursListeAlgorithme	Boucle « pour chaque algorithme »
Parcours de la liste des configurations à tester	UniteParcoursListeConfiguration	Boucle « pour chaque configuration »
Parcours de la liste des combinaisons à tester	UniteParcoursListeCombinaison	Boucle « pour chaque combinaison »
Parcours de la liste des répartitions à tester	UniteParcoursListeRepartition	Boucle « pour chaque répartition »
Génération d'un modèle de risque	GenererModele	Générer un modèle
Génération de statistiques descriptives en fonction de l'algorithme utilisé	GenererStatistiques	Générer des statistiques
Évaluation des performances du modèle (capacité à prédire le risque correctement)	ÉvaluerPerformances	Évaluer les performances
Combiner les modèles générés sur des répartitions différentes	CombinerModeles	Combiner les modèles partiels

Les résultats analysés dans le chapitre 5 ont été obtenus en déployant ce composant sur un nœud d'exécution physiquement situé dans les locaux de Télécom Bretagne à proximité de Brest (29) alors que, serveur web mis à part, les autres nœuds

d'exécution sont situés sur le site de l'Institut de cancérologie Gustave Roussy à Villejuif (94). Ce nœud d'exécution est une machine virtuelle*, dotée de 50 Go d'espace disque, qui est placée sur des disques durs tournant à 15 000 tours par minute (deux fois plus vite que celui d'un ordinateur personnel). Cette machine virtuelle peut accéder à 256 Go de mémoire vive et aux quatre processeurs AMD Opteron 6176 SE qui totalisent 48 cœurs physiques sur lesquels peuvent être répartis les fils d'exécution permettant de réaliser les fonctionnalités de génération des modèles de risque et de l'évaluation de leurs performances.

Dans ce chapitre, nous avons fait une proposition d'architecture fonctionnelle pour un système d'information qui permette de supporter la production et l'utilisation de scores de risque fondée sur les processus métiers modélisés dans le chapitre 3. Nous avons conçu l'architecture fonctionnelle du système d'information du risque de maladie alignée sur les processus métiers. Afin de pouvoir mener des expérimentations dans la génération de modèle de risque, nous avons proposé une architecture applicative du système informatique en mettant en évidence l'influence du métier sur le déploiement de ce système en particulier pour le composant applicatif chargé de la gestion des calculs.

Dans le chapitre suivant, nous utilisons ce composant applicatif au sein du système mis en place pour concevoir un modèle de risque pour le cancer du sein grâce à l'algorithme des plus proches voisins dans le contexte particulier d'une clinique de risque.

5 Estimation du risque de cancer du sein avec l'algorithme des plus proches voisins

La conception d'un score de risque qui soit adopté par un grand nombre d'utilisateurs qui n'ont pas de connaissance des méthodes de modélisation (médecins ou patients par exemple), implique d'utiliser une méthode de modélisation compréhensible. À cette fin, nous avons envisagé l'utilisation des arbres de décision et du concept des plus proches voisins.

Les premiers essais avec différents modes de construction d'arbres sous différentes configurations ne permettent d'obtenir un modèle d'évaluation performant que pour des arbres profonds et donc difficilement compréhensibles. Nous avons donc choisi de nous concentrer sur l'algorithme des plus proches voisins, mais son utilisation pose le problème de la prise en compte du contexte d'utilisation du score par rapport aux données médicales disponibles et le problème de la performance des scores produits par rapport aux performances de scores relevées dans la littérature.

En utilisant le processus proposé au chapitre 3 et grâce à l'architecture du système d'information détaillée au chapitre 4, nous présentons dans ce chapitre une adaptation de l'algorithme des plus proches voisins. Nous en évaluons les performances sur deux jeux de données : un jeu américain public pour permettre la comparaison des performances mesurées avec celles de la littérature et un jeu français pour pouvoir proposer un score de risque utilisable par les femmes françaises.

Concernant les résultats, nous obtenons de multiples modèles permettant d'évaluer le risque de cancer du sein et nous montrons que les performances sont comparables en termes de discrimination et supérieures en termes de calibration. Les experts du domaine choisissent une combinaison d'attributs composée des attributs âge de la femme, maladie bénigne, âge à la ménopause et nombre d'antécédents au premier degré de cancer du sein. Cette combinaison est à la fois adaptée au contexte d'utilisation et offre un haut degré de performance par rapport à la littérature.

5.1 Algorithme des plus proches voisins

Afin de répondre à l'exigence de compréhensibilité de la méthode de modélisation du risque de cancer du sein, nous avons fait le choix original d'utiliser l'algorithme des plus proches voisins pour prédire un niveau de risque dans un cadre de prévention en population générale. En effet, à notre connaissance, l'algorithme des plus proches

voisins n'a jamais été utilisé pour produire un score de risque dans le domaine de la prévention en santé. Pour éclairer les résultats qui sont présentés dans la suite du chapitre, nous expliquons dans cette partie le principe de fonctionnement général d'un tel algorithme. Nous expliquons notre choix d'utiliser une mesure de distance euclidienne pour déterminer la similarité entre les individus de la cohorte* et nous détaillons les résultats qui nous amènent à pondérer les voisins recrutés de manière uniforme lors de la construction du voisinage.

5.1.1 Principe de fonctionnement

Après une première présentation informelle du principe de l'algorithme des plus proches voisins et des problèmes qui se posent pour son utilisation, nous revenons sur l'origine de cette méthode et en faisons une description plus formelle. Enfin, nous expliquons la méthode retenue pour calculer un risque grâce à cette méthode.

5.1.1.1 Présentation de l'algorithme des plus proches voisins

Si nombre de méthodes de modélisation du risque, qu'elles soient paramétriques* ou non-paramétriques, se fondent sur le degré de similarité entre individus d'une base d'expériences pour évaluer le risque d'une personne extérieure à cette base, une méthode de type proches voisins a l'avantage d'avoir un fonctionnement simple et compréhensible, car calqué sur le mécanisme qu'utilise notre cerveau pour résoudre des situations déjà rencontrées dans le passé [Kolodner 84].

Dans le contexte d'une maladie, le principe d'un algorithme des plus proches voisins est d'utiliser la similarité d'une personne avec d'autres personnes choisies dans une base de connaissances passées. Ainsi, le risque de maladie est déterminé en fonction de la proportion d'individus similaires qui ont été atteints par cette maladie.

De cette nécessité de recruter des individus similaires pour estimer le niveau de risque d'une personne naissent plusieurs problématiques qui doivent trouver une réponse adaptée à notre problème de modélisation du risque de cancer du sein en population générale.

- Comment mesurer le degré de similarité entre le profil de la personne dont on souhaite évaluer le risque et le profil des personnes qui constituent notre base d'expériences ?
- Comment déduire le niveau de risque d'une personne à partir d'un ensemble de personnes similaires ?
- Combien de personnes similaires faut-il choisir pour obtenir une évaluation optimale du risque ?
- Comment considérer le voisinage de personnes similaires afin d'obtenir la meilleure performance de détection des profils à risque ?

5.1.1.2 Origine du concept des proches voisins

L'algorithme des plus proches voisins est un algorithme qui fait partie de l'apprentissage supervisé à base d'exemples. L'apprentissage est supervisé parce que les exemples utilisés sont déjà associés à une classe (*sain* ou *malade* dans notre cas) pour laquelle on cherche à déterminer la probabilité d'appartenance. On dit qu'il fonctionne à base d'exemples, car pour chaque nouvelle évaluation d'une probabilité d'appartenance à une classe, on revient aux exemples contenus dans une base d'expériences, à la différence des méthodes d'inductions qui permettent de produire des règles intermédiaires (par exemple les arbres de décision).

La méthode des plus proches voisins trouve ses origines dans la statistique : [Fix 51] avait analysé les propriétés d'une telle méthode et notamment sa consistance pour un nombre de voisins tendant vers l'infini, mais les premières utilisations qui ont été faites de cet algorithme étaient plus limitées. En effet, pour attribuer une classe à un exemple nouveau, elles se contentaient de chercher l'exemple le plus similaire de la base d'expériences pour en connaître la classe et l'attribuer au nouvel exemple [Johns 59, Sebestyen 62]. Des raffinements sont apparus par la suite quant aux moyens utilisés pour mesurer la similarité (voir le choix de la mesure de distance en partie 5.1.2.1, page 113), pour recruter les voisins (voir partie 5.1.3.1, page 117), pour traiter le voisinage de manière spécifique (voir partie 5.1.3.2, page 118) afin d'attribuer une classe en étudiant le comportement du taux d'erreur moyen [Cover 67].

5.1.1.3 Description formelle

Plus formellement, l'algorithme des plus proches voisins consiste en la recherche, dans un espace S de dimension D contenant un ensemble E de N points, des k points de E les plus similaires à x, x étant un point de S qui constitue l'exemple dont on cherche à évaluer la probabilité d'appartenance à une classe c. La figure 5.1 illustre le problème de la recherche de $k = 6$ voisins avec des points appartenants à deux classes dans un espace à deux dimensions centré sur x.

Dans notre travail de modélisation du risque de cancer du sein, D correspond au nombre de facteurs de risque utilisés pour donner la probabilité d'appartenance à la classe *malade*, la taille k du voisinage est déterminée en fonction de la répartition des classes dans les ensembles d'apprentissage ou de validation et est donc fonction de la prévalence du cancer du sein, N correspond à la taille de l'échantillon utilisé pour l'apprentissage et x correspond au profil de la femme dont on veut évaluer le risque.

5.1.1.4 Méthode de calcul du risque basé sur la prévalence

Une fois les distances calculées entre une personne dont il faut évaluer le risque et les individus de la base d'expériences, un voisinage de taille k est recruté. Pour

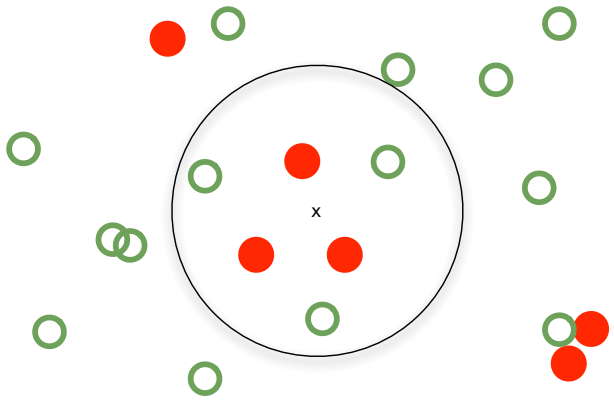

Figure 5.1 : Exemple de la recherche de plus proches voisins avec S centré sur x, $D = 2$, $N = 20$ et $k = 6$. La classe *sain* est symbolisée par des cercles verts et la classe *malade* est symbolisée par des disques rouges.

calculer le niveau de risque de la personne, nous avons choisi d'utiliser le concept de la prévalence* qui est maîtrisé par les médecins qui seront amenés à utiliser le score et facile à expliquer aux patientes. La prévalence se définit comme un nombre de cas d'une maladie dans une population donnée à un instant donné. Appliqué au voisinage recruté et au cancer du sein, le niveau de risque évalué est donc le rapport du nombre de personnes affectées par le cancer du sein sur la taille du voisinage.

Pour utiliser le principe de similarité permettant de constituer un voisinage d'individus semblables, il faut répondre à deux questions. La première est de savoir quelle mesure utiliser pour déterminer la similarité entre l'individu dont on veut calculer le risque et chacun des individus du jeu d'apprentissage. La seconde est de savoir comment le voisinage est construit et traité.

5.1.2 Distance d'un profil à son voisin

Il existe diverses manières de mesurer la similarité entre deux éléments. Pour faciliter la compréhension de la méthode de construction du score de risque et afin de tester la performance d'une méthode éprouvée, nous avons choisi d'utiliser une mesure de distance pour calculer la similarité entre deux profils de femmes.

Dans cette sous-partie, nous décrivons le processus qui a amené au choix d'une distance euclidienne non-pondérée comme mesure de similarité entre deux individus, plus la distance étant faible, plus les individus étant considérés comme similaires. Ce choix a été fait après avoir testé différentes déclinaisons de la distance de Minkowski et la possibilité de pondérer les éléments dans le calcul de la distance.

5.1.2.1 Choix de la mesure

Pour mesurer la similarité entre deux individus de notre base d'apprentissage, afin de ne retenir que les k individus les plus ressemblants au profil pour lequel on souhaite calculer un risque, nous avons envisagé les différentes déclinaisons d'une distance de Minkowski qui permet de mesurer une distance d entre deux femmes représentées par des points X et Y de coordonnées $(x_1, x_2, \ldots, x_n)$ et $(y_1, y_2, \ldots, y_n)$:

$$d_p(X,Y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}\,, \tag{5.1}$$

où n est le nombre de coordonnées pour chaque point et p la puissance retenue pour calculer la distance.

Bien que la distance habituellement retenue pour l'algorithme des plus proches voisins est la distance euclidienne avec $p = 2$ [Fix 51], nous avons choisi de tester quatre valeurs différentes de $p = 1$ à $p = 4$ pour déterminer quelle est la meilleure mesure de distance à utiliser avec les données à disposition. Nous avons choisi de limiter le nombre de valeurs testées pour p, car lorsque p tend vers l'infini, la nature de la distance change pour ne conserver que la dimension de plus grande distance tel que :

$$d_\infty(X,Y) = \max(|x_i - y_i|) \qquad \text{avec } i = 1, \ldots, n\,. \tag{5.2}$$

Afin de mesurer l'influence de la valeur de p sur les performances, nous avons choisi de tester toutes les combinaisons possibles C_n^t de taille $t = 2$ et $t = 3$ attributs parmi les $n = 8$ attributs retenus en priorité par l'expert du domaine sur le jeu issu de l'étude E3N, soit $\sum_{t=2}^{3} C_n^t = \sum_{t=2}^{3} \frac{n!}{t!(n-t)!} = 84$ combinaisons. Nous n'avons pas retenu les combinaisons de taille 1 en raison du faible nombre de catégories de score produites avec seulement un attribut. En effet, avec peu d'attributs il y a peu de combinaisons de modalités d'attributs et par conséquent, peu de risques différents à attribuer. En conséquence, peu de points peuvent êtres utilisés pour tracer la courbe ROC afin de mesurer la discrimination ce qui conduit à une courbe ROC peu précise et donc une mesure de performance peu fiable. Nous nous sommes limités à trois attributs maximum par combinaison pour des raisons de temps de calcul. Nous avons choisi de tester différentes tailles de voisinage : de 1 000 à 15 000 voisins par palier de 1 000 voisins, soit 15 tailles de voisinage différentes sur chacune des quatre répartitions de validation croisée définies au départ. Pour chaque valeur de p, ce sont donc $84 \times 15 \times 4 = 5\,040$ scores qui ont été générés et dont la performance a été mesurée.

Puisque seules les combinaisons à deux ou trois attributs ont été utilisées, les performances moyennes et maximales présentées ne sont pas représentatives des performances globales présentées dans la suite du chapitre.

Le tableau 5.1 présente les résultats obtenus en termes de discrimination avec l'AUC (aire sous la courbe ROC) et de calibration avec le rapport E/O du nombre

Tableau 5.1 : Discrimination (mesurée par l'AUC) et calibration (mesurée par le rapport E/O) par type de distance de Minkowski sur le jeu de données issu de l'étude E3N.

	$p = 1$	$p = 2$	$p = 3$	$p = 4$
AUC moyenne	0,566	0,566	0,566	0,566
AUC médiane	0,564	0,565	0,565	0,565
Meilleure AUC	0,616	0,618	0,617	0,617
E/O moyen	0,970	0,972	0,972	0,972
E/O médian	0,972	0,974	0,974	0,974
Meilleur E/O	1,000	1,000	1,000	1,000

estimé de cas de cancer du sein (classe positive ou *malade*) sur le nombre observé de cas de cancer du sein.

On observe que les performances moyennes sont inférieures avec $p = 1$, cette distance est donc éliminée. On observe ensuite que les performances sont stables ou inférieures quand la valeur de p augmente. Globalement, les performances avec d'autres valeurs de p que la valeur naturelle 2 généralement utilisée dans le calcul de ce type de distance, ne permettent pas d'améliorer les performances. Nous choisissons donc d'utiliser la distance avec la valeur de p la plus performante, c'est de plus une distance simple et connue de tous : nous utilisons la distance de Minkowski avec $p = 2$ dans la suite de ce chapitre.

5.1.2.2 Pondération de la mesure

La distance de Minkowski est la somme des différences entre deux valeurs d'attributs, la distance globale est donc la somme de la distance entre deux points dans chaque dimension. Son utilisation, quelle que soit la valeur choisie pour p, implique que toutes les dimensions contribuent de manière équivalente à la distance totale.

En effet, toutes les mesures de performance présentées dans ce chapitre sont effectuées grâce à des attributs dont la valeur a été réduite afin d'éviter qu'un attribut ne domine un autre dans le calcul de distance. Cette transformation est effectuée en divisant les valeurs utilisées par l'écart-type (voir [Saporta 06], page 168). Une normalisation complète des données ajoutant le centrage des données à la réduction n'est pas nécessaire, vu la formule de distance utilisée : en effet, la soustraction dans la distance euclidienne donne le même résultat que les valeurs aient été centrées ou non.

Si toutes les dimensions contribuent de manière équivalente à la distance totale, tous les attributs utilisés pour décrire un individu ne participent pas forcément de manière égale à la similarité entre eux. Appliqué à la maladie, cela signifie que tous les facteurs de risque n'ont pas un impact identique sur le risque de cancer du sein.

Tout en gardant l'objectif d'utiliser une mesure de similarité qui reste compréhensible par les futurs utilisateurs, nous avons testé l'impact d'une pondération de chaque attribut utilisé dans la mesure de distance. L'objectif étant de rapprocher les voisins en fonction des dimensions qui contribuent le plus à élever le niveau de risque.

Nous avons choisi de comparer l'influence de trois pondérations différentes sur la performance mesurée.

La première méthode consiste à ne pas pondérer la distance. Les deux autres méthodes retenues pour fixer les coefficients de pondération font appel à la connaissance des experts du domaine par l'intermédiaire de la littérature du domaine. En utilisant une suite de méta-analyses* réalisées par des épidémiologistes sur le cancer du sein appelée méta-analyse d'Oxford (voir par exemple [Group 96, Group 97, Group 02]), nous avons extrait les risques relatifs* calculés par les épidémiologistes pour les attributs que nous possédons.

La première méthode consiste à utiliser un risque relatif moyen par attribut tandis que la seconde méthode consiste à détailler le coefficient de pondération attribué en fonction de la modalité de l'attribut dans le calcul de similarité.

Tableau 5.2 : Coefficients utilisés pour pondérer la distance selon la dimension *Nombre de parents atteints au premier degré.*

Parents atteints	Pondération
0	1,0
1	1,8
2	2,9
3 et plus	3,9

À titre d'exemple, le tableau 5.2 présente les pondérations utilisées lorsque l'attribut *Nombre de parents atteints au premier degré* entre en compte dans le calcul de la distance suivant la seconde méthode. Chaque valeur de risque relatif correspond à un nombre de parents atteints par le cancer du sein. Le risque relatif correspond à l'augmentation du risque de cancer du sein calculé par les épidémiologistes dans la méta-analyse d'Oxford qui regroupe plusieurs analyses épidémiologiques.

Pour prendre en compte la pondération et en fixant $p = 2$, la formule de distance 5.1, page 113, devient donc :

$$d_2(X,Y) = \sqrt[2]{\sum_{i=1}^{n} \omega_i \, |x_i - y_i|^2}\,, \tag{5.3}$$

avec ω_i correspondant au coefficient de pondération issu de la littérature épidémiologique.

Les résultats sur le jeu de données constitué à partir de la base E3N sont regroupés dans le tableau 5.3. On observe que l'utilisation de pondérations issues de

Tableau 5.3 : Discrimination (mesurée par l'AUC) et calibration (mesurée par le rapport E/O) en fonction de la méthode de pondération sur le jeu de données issu de l'étude E3N.

	Pondération uniforme	Pondération simple	Pondération détaillée
AUC moyenne	0,566	0,549	0,550
AUC médiane	0,565	0,552	0,552
Meilleure AUC	0,618	0,614	0,618
E/O moyen	0,972	0,983	0,985
E/O médian	0,974	0,992	0,992
Meilleur E/O	1,000	1,000	1,000

la littérature épidémiologique et calculées sur une population internationale dégrade les performances de discrimination du score de risque construit et testé sur une population française avec une pondération uniforme. Les performances sont moins dégradées avec la méthode de pondération détaillée (qui est plus fine, car moins dépendante des modalités) que la première : en effet, la pondération est affinée en fonction du nombre de parents atteints avec la seconde méthode alors que la première méthode conduit à utiliser la même pondération, quel que soit le nombre de parents.

La calibration mesurée avec le rapport E/O est légèrement améliorée avec les pondérations simple ou détaillée, mais reste excellente avec une pondération uniforme avec un rapport moyen supérieur à 0,97 quand les modèles évoqués au chapitre 2 affichent un rapport compris entre 0,5 et 0,8 [Gail 89]. Pour effectuer notre choix de pondération, nous choisissons donc de privilégier le net gain en termes de discrimination plutôt que le faible gain en termes de calibration.

La dégradation des performances de discrimination par rapport à la pondération uniforme peut être expliquée par la différence entre la population utilisée pour construire notre score de risque et les populations utilisées pour calculer les risques relatifs de maladie dans la méta-analyse d'Oxford. De plus, nous choisissons de ne pas tenter d'optimiser la valeur des pondérations pour mieux modéliser le risque : une telle démarche pourrait en effet conduire à sur-apprendre sur le jeu de données utilisé pour réaliser les tests, l'augmentation des performances mesurées serait alors artificielle et ne correspondrait pas à la performance réelle obtenue sur la population générale dans laquelle sera utilisé le score.

Nous choisissons donc d'utiliser le système de pondération uniforme qui permet d'atteindre les meilleures performances sans surapprentissage sur les données.

5.1.3 Construction du voisinage

La construction du voisinage est une étape cruciale dans l'utilisation de l'algorithme des plus proches voisins. Le niveau de risque étant calculé sur le voisinage, la qualité de la prédiction dépend des voisins recrutés et de la manière dont ils sont considérés en fonction de leur éloignement au profil de la femme dont on veut évaluer le risque.

Dans cette partie, nous expliquons la manière dont nous fixons la taille du voisinage par l'intermédiaire d'un mode de recrutement des voisins indépendant de leur classe et la manière dont nous considérons le voisinage recruté en fonction de la distance au profil de la femme dont on veut évaluer le risque.

5.1.3.1 Type de recrutement des voisins

L'utilisation de l'algorithme des plus proches voisins se caractérise par l'importance d'un paramètre, celui du nombre de voisins recrutés pour constituer le voisinage.

Il est important de noter que les voisinages construits pour chaque individu dont on veut mesurer le risque n'ont pas tous une taille qui vaut strictement k, mais que k constitue la taille minimum du voisinage. En effet, plusieurs individus du jeu de données peuvent avoir des valeurs d'attributs identiques, ce qui implique une distance égale par rapport à l'individu dont on veut mesurer le risque. En conséquence, la taille réelle du voisinage construit peut être supérieure à la taille définie pour k qui n'est qu'une valeur minimum.

Soit k le nombre de voisins minimum inclus dans le voisinage. Il existe deux façons d'ajouter des voisins au voisinage : la première est de considérer les k_v voisins indépendamment de leur classe (*sain* ou *malade* dans notre cas). Mais notre classe *malade* étant très peu représentée dans les données, nous souhaitons non seulement mesurer la performance d'un algorithme dont le voisinage est construit en recrutant k_v voisins indépendamment de leur classe, mais également mesurer sa performance en considérant k_m, non pas comme la taille minimum totale du voisinage, mais comme le nombre d'individus de la classe *malade* compris dans le voisinage. Dans ce dernier cas, l'objectif est de s'assurer de la présence d'un nombre minimum d'individus malades dans le voisinage construit.

Nous testons donc les performances obtenues en fonction de deux manières de recruter les voisins : une fois la distance mesurée entre l'individu dont on veut mesurer le risque et les autres individus du jeu de données, le voisinage est construit en ajoutant successivement au voisinage en construction, tous les voisins situés à une distance d, tant que k n'est pas atteint, soit en termes d'individus (k_v voisins indépendamment de leur classe), soit en termes d'individus ayant la classe *malade* (k_m voisins malades).

Le tableau 5.4 présente les performances mesurées en fonction de la technique de recrutement utilisée pour construire le voisinage nécessaire au calcul du score. Une

Tableau 5.4 : Discrimination (mesurée par l'AUC) et calibration (mesurée par le rapport E/O) en fonction du mode de recrutement du voisinage sur le jeu de données issu de l'étude E3N.

	k voisins	k malades
AUC moyenne	0,551	0,545
AUC médiane	0,555	0,549
Meilleure AUC	0,618	0,598
E/O moyen	0,991	0,988
E/O médian	0,998	0,996
Meilleur E/O	1,000	1,000

distance non-pondérée de Minkowski avec $p = 2$ a été utilisée sur le jeu de données construit à partir des données de l'étude E3N.

On constate que la performance, exprimée sous la forme d'une AUC (aire sous la courbe ROC) ou de rapport nombre estimé de cas sur nombre observé E/O, diminue légèrement en définissant le voisinage en fonction du nombre de malades recrutés plutôt qu'en fonction du nombre total d'individus recrutés. Au vu de cette performance inférieure, nous choisissons de recruter les plus proches voisins indépendamment de la classe des individus.

Dans notre contexte d'utilisation, l'algorithme des plus proches voisins n'est donc pas affecté par le déséquilibre de la répartition des classes dans les données que nous utilisons.

5.1.3.2 Poids des voisins dans le voisinage

Si les performances ne sont pas améliorées en fonction du type de recrutement des voisins, nous avons vérifié si la diminution du poids de chaque voisin en fonction de sa distance à l'individu dont on veut mesurer le risque, permet d'augmenter les performances [Dudani 76]. En effet jusqu'à présent, quelle que soit la position des voisins dans le voisinage, ils participent tous de manière équivalente au calcul du score (rapport du nombre d'individus de la classe positive sur le nombre total d'individus).

Pour mesurer l'impact d'une telle pondération du poids de chaque voisin en fonction de la distance à l'individu dont on veut évaluer le risque, nous avons testé l'influence de plusieurs fonctions pour faire diminuer le poids des voisins en fonction de leur position dans le voisinage, et donc dans le score calculé. Plus un voisin est situé loin de l'individu dont on souhaite évaluer le risque, moins il influera le calcul du score.

La figure 5.2 présente les fonctions utilisées pour faire décroître la pondération appliquée à chaque voisin en fonction de sa distance, notée d, au profil à évaluer sachant que le $k$ième voisin est à une distance notée d_{max}. Sur le graphique de gauche,

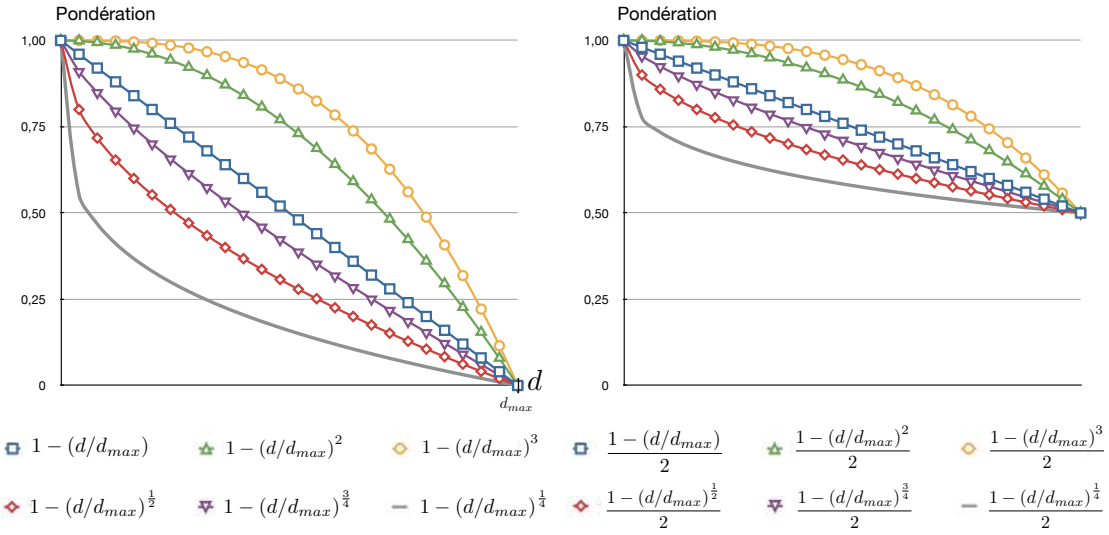

Figure 5.2 : Fonctions de décroissance de la pondération appliquées aux voisins en fonction de leurs distances à l'individu à évaluer.

les pondérations diminuent de 1,0 (pour un voisin à une distance 0 de l'individu dont on veut évaluer le risque) à 0 pour un individu situé à la distance maximale d_{max} de l'individu dont on veut évaluer le risque. Cela signifie qu'un voisin à la distance d_{max} ne rentre pas en compte dans le calcul de la prévalence qui constitue la valeur du score produit. Sur le graphique de droite, la pondération minimale pour un voisin à la distance d_{max} vaut 0,5 pour que tous les voisins recrutés influent sur le score calculé.

Le tableau 5.5 présente les résultats obtenus pour les douze fonctions de pondération testées.

La colonne P^0 correspond aux résultats obtenus sans application d'une fonction ou, plus précisément, à l'application d'une pondération de 1,0 à tous les voisins recrutés.

Quelle que soit la fonction de pondération utilisée, on note une très légère amélioration des performances moyennes et médianes. Cette augmentation est logique si l'on considère que l'augmentation de la taille du voisinage consiste à inclure plus de voisins éloignés qui ressemblent moins à l'individu dont on veut estimer le risque. Or, la pondération diminue l'importance des voisins les plus éloignés, ce qui atténue la baisse de performance due à l'augmentation de la taille du voisinage. En conséquence, les performances moyennes et médianes augmentent par rapport à une mesure de performance effectuée sans pondérer les voisins en fonction de la distance.

En revanche, le pic de performance n'augmente pas. Pour la calibration (rapport nombre estimé de cas sur nombre observé), le maximum est déjà atteint avant utili-

Tableau 5.5 : Discrimination et calibration en fonction de la pondération appliquée aux voisins avec $P = 1 - (d/d_{max})$ sur le jeu de données issu de l'étude E3N.

	P^0	$P^{\frac{1}{4}}$	$P^{\frac{1}{2}}$	$P^{\frac{3}{4}}$	P	P^2	P^3
AUC moyenne	0,551	0,553	0,553	0,552	0,552	0,552	0,552
AUC médiane	0,555	0,558	0,557	0,557	0,556	0,556	0,556
Meilleure AUC	0,618	0,618	0,618	0,618	0,618	0,618	0,618
E/O moyen	0,991	0,997	0,996	0,996	0,996	0,995	0,994
E/O médian	0,998	1,001	1,001	1,001	1,001	1,000	1,000
Meilleur E/O	1,000	1,000	1,000	1,000	1,000	1,000	1,000
		$P^{\frac{1}{4}}/2$	$P^{\frac{1}{2}}/2$	$P^{\frac{3}{4}}/2$	$P/2$	$P^2/2$	$P^3/2$
AUC moyenne		0,552	0,552	0,552	0,552	0,552	0,552
AUC médiane		0,556	0,556	0,556	0,556	0,556	0,556
Meilleure AUC		0,618	0,618	0,618	0,618	0,618	0,618
E/O moyen		0,995	0,995	0,994	0,994	0,994	0,994
E/O médian		1,000	1,000	1,000	1,000	1,000	1,000
Meilleur E/O		1,000	1,000	1,000	1,000	1,000	1,000

sation du système de pondération. Pour la discrimination, la meilleure AUC obtenue est de 0,618 avec ou sans pondération. Or, puisque parmi les différentes tailles de voisinage d'une même combinaison d'attributs, c'est le score pour lequel k génère la plus haute valeur d'AUC qui sera présenté aux experts du domaine, l'utilisation d'un tel système de pondération n'apportant pas de gain sur le pic de performance, nous choisissons de ne pas l'utiliser.

Dans cette partie, nous avons montré quels paramétrages sont les plus adaptés à notre contexte et au type de données que nous utilisons. Nous avons choisi de construire un score de risque en utilisant une distance de Minkowski pour laquelle $p = 2$ avec une pondération uniforme quel que soit l'attribut utilisé et sa valeur. Les voisins sont recrutés indépendamment de leur classe et chaque voisin possède un poids égal dans le voisinage recruté.

5.2 Préparation du jeu de données

Après la présentation de l'algorithme des plus proches voisins dont nous avons détaillé la configuration dans la partie 5.1, nous détaillons les deux jeux de données que nous avons choisi pour modéliser le risque de cancer du sein. Nous utilisons une base américaine que l'on nomme BCSC, pour « Breast Cancer Surveillance Consortium », du nom du consortium américain chargé de rassembler les données puis d'en publier une partie. Puis, nous utilisons une base française que l'on nomme E3N, pour « Étude Épidémiologique auprès de femmes de la MGEN » (Mutuelle Générale de l'Éducation Nationale). Enfin, nous décrivons les choix faits pour la construction des jeux de données : ces choix doivent permettre une comparaison aisée avec les scores existant pour les données américaines et une manipulation aisée de l'outil de calcul du score de risque par les utilisateurs.

5.2.1 Données du BCSC

Les données du « Breast Cancer Surveillance Consortium » ne sont pas entièrement disponibles publiquement, seul le jeu de données utilisé par [Barlow 06] peut être téléchargé en ligne [Consortium 06]. Afin de pouvoir comparer nos résultats, nous avons choisi de le conserver aussi proche que possible de l'original. Ce jeu de données est constitué à partir des réponses de femmes venues dans un centre de dépistage du cancer du sein pour passer une mammographie. Dans cette partie nous décrivons les attributs disponibles et les spécificités de leur distribution.

5.2.1.1 Attributs

Le jeu de données disponible en ligne est constitué de 12 attributs qui décrivent le profil d'une femme au moment de la mammographie quand les informations sont recueillies et deux attributs qui renseignent sur le statut de la femme par rapport au cancer dans les 12 mois qui suivent la mammographie. Parmi la douzaine d'attributs décrivant le profil de la femme qui a subi la mammographie, deux caractérisent son origine ethnique et ne sont pas retenus par l'épidémiologiste qui produit une liste filtrée contenant les dix autres attributs.

Le tableau 5.6 regroupe les dix attributs de liste filtrée par l'épidémiologiste et validée par l'analyste comme le prévoit le processus présenté en chapitre 3. Il est intéressant de noter que les valeurs continues des attributs ont été discrétisées, c'est le cas pour l'âge de la femme, l'âge de la femme à la première naissance et l'indice de masse corporelle. L'avantage est un nombre de modalités moins important à traiter par les outils de modélisation, l'inconvénient qui en résulte est la possible perte d'information, notamment sur l'âge qui est un attribut fortement prédictif dans le cas du cancer du sein.

Étant donné que les données proviennent de différents centres de dépistage pour le cancer du sein aux États-Unis et que ces derniers n'utilisent pas de questionnaire

Tableau 5.6 : Tableau des attributs du jeu BCSC filtré et validé, leurs modalités et la proportion de valeurs manquantes.

Attribut	Nom abrégé	Modalités	Valeurs manquantes
Statut ménopausique	statmeno	{Préménopause, postménopause}	7,6 %
Âge	age	10 catégories de 35 à 84 ans	0 %
Densité mammaire	dens	4 catégories	26,3 %
IMC	imc	4 catégories de 10 à plus de 35	55,9 %
Âge à la 1ère naissance	agenai	{Avant 30 ans, après 30 ans}	55,5 %
Ancédent au 1er degré	kdeg1	{Aucun, un, deux ou plus}	15,2 %
Biopsie précédente	biop	{Oui, non}	10,5 %
Résultat mammographie précédente	precmamm	{Négatif, faux positif}	23,4 %
Type ménopause	typemeno	{Naturelle, chirurgicale, inconnu}	52,1 %
Traitement hormonal substitutif de la ménopause	ths	{En cours, non}	41 %
Cancer du sein	ks	{Oui, non}	0 %

standardisé, le taux de valeurs manquantes est différent selon les attributs. Ce taux est le plus important pour l'indice de masse corporelle ajouté aux questionnaires en cours de recueil. Il est également très élevé pour l'âge de la femme lors de la naissance du premier enfant. Pour le type de ménopause, le taux de valeurs manquantes élevé s'explique par le codage de l'attribut. En effet, le jeu de données mis à disposition du public mélange le fait que l'information ne soit pas possédée et le fait que la femme ne soit pas encore ménopausée.

5.2.1.2 Distribution des valeurs

La figure 5.3 présente la distribution des groupes d'âge dans les données. Celles-ci sont basées sur une population qui se rend dans des centres de dépistage aux États-Unis, il est donc logique d'observer une faible représentation des femmes de moins de 40 ans qui ne sont pas encouragées à réaliser de dépistage dans ce pays. La majorité des femmes est représentée dans les groupes d'âge de 40 à 60 ans qui sont la cible des campagnes de prévention. Puis, logiquement, en même temps que l'âge augmente, la représentation des groupes d'âges diminue à cause d'une moindre

fréquentation des centres de dépistage.

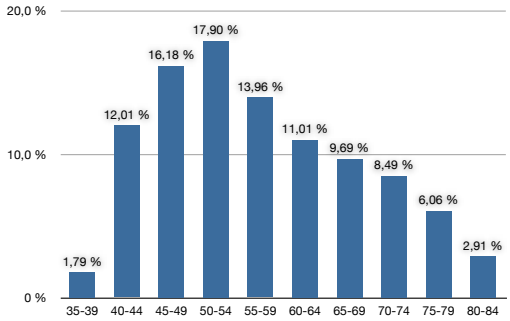

Figure 5.3 : Distribution des groupes d'âge au sein du jeu de données BCSC.

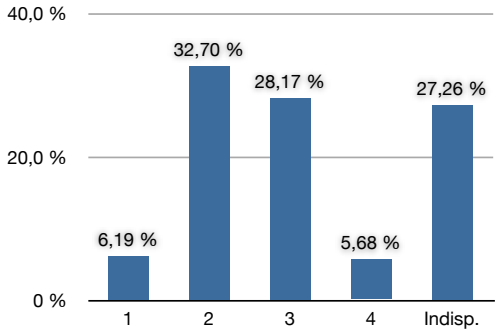

Figure 5.4 : Distribution des niveaux de densité mammaire au sein du jeu de données BCSC.

Il est important que la distribution des groupes d'âge reflète la réalité d'une manière acceptable, car l'âge est un facteur de risque parmi les plus importants (voir partie 1.2.3, page 17).

La densité mammaire est habituellement un bon marqueur du risque de cancer du sein : le système d'évaluation utilisé (BI-RADS, [Reston 03]) est construit de telle sorte que le radiologue a le choix entre quatre groupes différents pour chaque femme : sein très dense (niveau 4), moyennement dense (niveau 3), avec des opacités fibroglandulaires dispersées (niveau 2) ou entièrement graisseux (niveau 1). Dans les faits, voir tableau 5.4, les radiologues utilisent en majorité les deux niveaux intermédiaires qui représentent environ 84 % des mammographies pour lesquelles on détient l'information (60,87 % des 72,71 % données disponibles). De plus, l'attribution des niveaux est réalisée par des personnes différentes selon les centres de dépistage et sur une longue période de temps peu propice à permettre une reproductibilité satisfaisante des classements. Ces faiblesses conduisent à l'utilisation d'une information dégradée qui pourrait être meilleure à l'avenir avec la généralisation d'outils de radiographie capables de délivrer des marqueurs de densité plus nombreux, de meilleure qualité et de manière reproductible dans le temps.

Parmi les attributs habituellement les plus prédictifs que l'on retrouve dans les scores de risque de cancer du sein, on trouve également les antécédents familiaux de cancer du sein au premier degré (mère, soeur, fille), le statut ménopausique, la prise d'un traitement hormonal substitutif à la ménopause et l'âge de la femme à la naissance du premier enfant. La distribution des valeurs de ces attributs, présentée dans la figure 5.5, est conforme à la connaissance des épidémiologistes avec des modalités très déséquilibrées pour certains d'entre eux. C'est le cas pour les antécédents familiaux de cancer : pour la majorité des réponses recueillies, aucun antécédent n'est déclaré, un faible nombre avec un antécédent est déclaré (environ 12 %) et un très faible nombre de personnes déclare deux antécédents au premier degré (0,65 %) ce

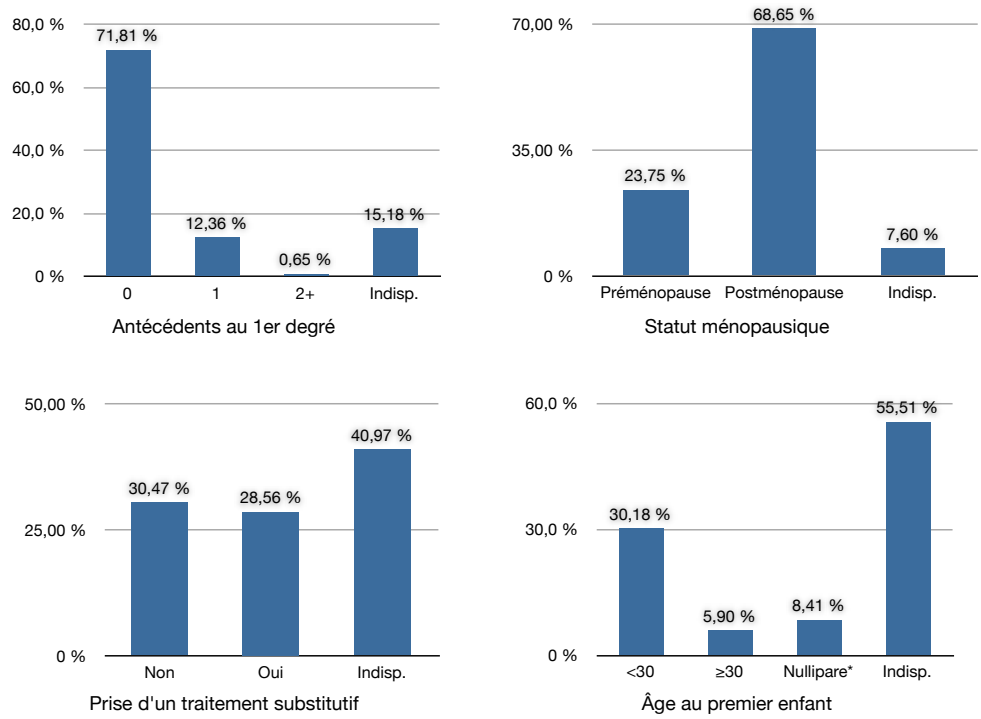

Figure 5.5 : Distribution des antécédents de cancer du sein au premier degré, du statut ménopausique, de la prise d'un traitement substitutif après la ménopause et de l'âge de la femme à son premier enfant, au sein du jeu de données BCSC.

qui devrait permettre de prendre en compte une partie des causes génétiques du cancer du sein chez les familles les plus à risque dans notre score.

Enfin, en ce qui concerne la proportion de cas de cancer recensée dans les douze mois qui suivent le recueil des informations, on constate un très fort déséquilibre comme discuté en partie 3.2.2, page 68.

Pour ce jeu de données, la distribution est très inégale avec 9 314 cas de cancer repertoriés parmi les 2,4 millions d'observations réalisées (figure 5.6). En termes d'apprentissage, cela signifie que la classe positive, dont il faut apprendre les caractéristiques pour créer le score de risque, ne représente que 0,39 % des données disponibles.

Dans cette partie, nous avons présenté les données issues d'une base de données américaines publiques que nous utiliserons pour modéliser le risque de cancer du sein afin de comparer les performances de prédiction avec celles de [Barlow 06]. Nous avons mis en évidence un taux de valeurs manquantes élevé pour certains attributs dont l'indice de masse corporelle et l'âge de la femme à la naissance de son premier enfant. Nous avons souligné la faiblesse de la méthode utilisée pour mesurer la densité mammaire et nous avons caractérisé le déséquilibre des données utilisées.

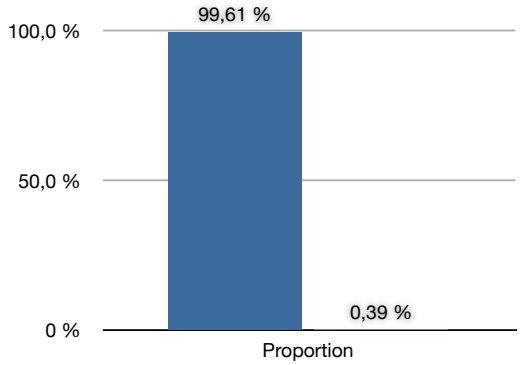

Figure 5.6 : Distribution des cas de cancer, repertoriés dans les douze mois après recueil des informations, au sein du jeu de données BCSC.

5.2.2 Données de la cohorte E3N

Dans le but d'utiliser le score de risque sur une population de femmes françaises, nous avons choisi d'utiliser la base de données construite pour l'étude E3N (décrite en partie 3.2.1.2, page 66) grâce aux réponses d'environ 100 000 femmes françaises issues du fichier d'une mutuelle nationale à destination des personnels de l'Éducation Nationale. Le mode de sélection des attributs, la distribution des attributs choisis et les choix effectués concernant la fenêtre de temps, de validation croisée et gestion des valeurs manquantes sont décrits dans cette partie en suivant le processus décrit au chapitre 3.

5.2.2.1 Attributs

Au contraire du jeu limité fourni en ligne par le BCSC, nous avons accès à toutes les données de l'étude E3N. Parmi les données disponibles, on trouve :

- les données brutes issues de la numérisation des questionnaires papier renvoyés par les femmes qui participent à l'étude,
- les données nettoyées en fonction d'une validation visuelle des questionnaires numérisés, par exemple pour corriger les erreurs dues à une mauvaise lecture optique du logiciel de reconnaissance,
- les données dites *générées* qui sont construites pour faciliter le travail des épidémiologistes, c'est par exemple le cas de l'attribut cible, le fait d'être atteint d'un cancer du sein ou non : la donnée utilisée n'est ni la donnée brute déclarée sur le questionnaire, ni la donnée nettoyée, mais une donnée construite à partir des déclarations des femmes sur plusieurs années, validée pour chacun des milliers de cas de cancer du sein auprès du médecin, afin de générer un historique précis de l'histoire de chaque participante sur ce point. C'est également le cas des traitements de la ménopause qui sont générés à partir

des déclarations des femmes, d'éventuelles données de remboursement de la mutuelle. La quantité d'alcool consommée est également une donnée générée, mais pour celle-ci à partir des questionnaires alimentaires, l'alcool pouvant se trouver dans différents aliments.

Choix des attributs : La recherche des facteurs de risque pour le cancer du sein a conduit les épidémiologistes de l'étude E3N à tester de manière systématique toutes les données à disposition pour mesurer l'impact des facteurs de risque qu'elles représentent sur le cancer du sein. Par conséquent, nous choisissons de restreindre le nombre de facteurs de risque considérés pour construire notre score de risque et d'éliminer une stratégie habituellement utilisée en fouille de données qui consiste à tirer parti de la puissance de calcul pour tester toutes les solutions de manière systématique. De plus, ce type de stratégie a le défaut de ne pas utiliser la connaissance des experts, auxquels nous avons accès, sur les données et sur le cancer du sein. Enfin, au vu de la quantité d'attributs à tester et du choix d'un algorithme compréhensible, mais consommateur de ressources, comme celui des proches voisins, les temps de calcul auraient été considérablement augmentés.

Tableau 5.7 : Attributs choisis en fonction de leur disponibilité dans E3N et de la littérature épidémiologique du cancer du sein.

Âge	
Maladies bénignes du sein	Nombre d'enfants
Âge à la première naissance	Statut tabagique
IMC	Parent atteint au premier degré
Durée allaitement	Nombre d'avortements
Âge à la ménarche*	Âge à la ménopause
Type de ménopause	Alcool
Biopsie	Traitement de la ménopause

En fonction de la littérature épidémiologique sur le cancer du sein (décrite en partie 1.2.3, page 17), et en suivant le processus métier (décrit page 76) à la base de l'architecture proposée (décrite page 89), un premier jeu de quinze attributs est constitué par l'*expert en fouille de données* (voir tableau 5.7).

Conformément au processus, la liste des attributs doit être filtrée par un épidémiologiste : seuls huit attributs sont validés par l'analyste en fonction de leur facilité d'utilisation dans le contexte d'une clinique du risque d'une part et en fonction de leur impact connu sur le risque de cancer du sein dans la littérature d'autre part. Les attributs comme le nombre d'enfants, la durée de l'allaitement, le nombre d'avortements ou l'indice de masse corporelle ne sont pas retenus à cause de leur faible

impact connu sur le risque de cancer du sein. Le statut tabagique est éliminé, car l'attribut est en cours de nettoyage et n'est donc pas disponible.

L'attribut décrivant le traitement hormonal de la ménopause n'est pas retenu à cause de l'évolution des prescriptions ces dernières années. En effet, leur consommation a chuté depuis que des travaux ont montré le faible bénéfice de la prise de tels traitements [Fournier 05] comparé au risque de cancer du sein. Les femmes qui utiliseront le score de risque que nous construisons auront des consommations de traitements hormonaux trop différentes de celles contenues dans les données E3N utilisées. L'épidémiologiste élimine donc cet attribut.

Cette liste d'attributs est ensuite validée par l'analyste en fonction des besoins des utilisateurs et donc, en fonction du contexte d'utilisation. Les attributs utilisés pour construire le score de risque sont présentés dans le tableau 5.8.

Tableau 5.8 : Tableau des attributs utilisés et leurs modalités.

Attribut	Nom abrégré	Modalités
Âge	age	Arrondi à l'année près de 46 à 72 ans
Âge à la 1ère naissance	agenai	{0-20 ; 21-25 ; 26-30 ; 30+} en années
Âge à la ménopause	agemeno	Arrondi à l'année près de 40 à 70 ans
Type ménopause	typemeno	{Naturelle, chirurgicale, inconnu, non-ménopausée}
Âge à la ménarche	menarche	{0-9 ; 10-11 ; 12-13 ; 14 et plus}
Déclaration d'une biopsie	biop	{Oui, non}
Antécédent au 1er degré	kdeg1	{0 ; 1 ; 2 ; 3 ; 4 et plus}
Déclaration d'une maladie bénigne du sein	mbs	{Oui, non}
Cancer du sein	ks	{Oui, non}

Grâce à la répétition des questionnaires tous les deux à trois ans, les taux de valeurs manquantes pour les attributs sont inférieurs à 5 %, ils sont donc faibles si on les compare à ceux du jeu de données du BCSC. Dans le jeu que nous construisons pour E3N, lorsqu'une valeur est manquante, nous imputons par la médiane si la valeur ne peut pas être remplacée par une valeur approchée obtenue à l'occasion d'un autre questionnaire que celui retenu. Par exemple, l'âge aux premières règles ou l'âge à la première naissance ne doit pas avoir évolué au fur et à mesure des questionnaires.

Choix de la fenêtre de temps : Comme indiqué en partie 3.2.1.2, page 66, tous les deux à trois ans, des questionnaires sont envoyés aux femmes qui participent à l'étude. Pour un même attribut, il peut donc y avoir plusieurs valeurs disponibles selon le questionnaire considéré. Afin de construire le score de risque, nous avons choisi de fixer une fenêtre de temps. Le début de la fenêtre de temps correspond à l'instant où les valeurs des attributs sont fixées, dans notre cas en 1997 au questionnaire Q5 à partir duquel un maximum d'attributs est disponible. Nous avons choisi de construire un score de risque à l'échéance de 10 ans.

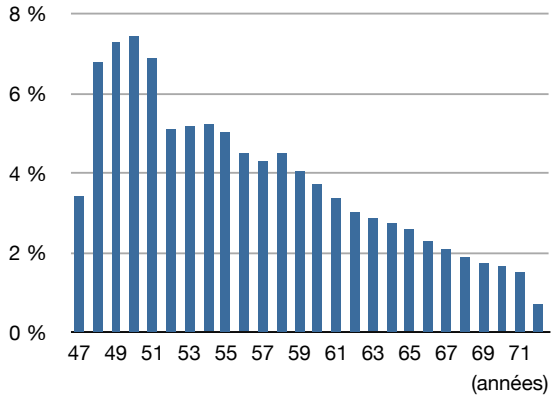

Figure 5.7 : Distribution de l'âge des femmes au sein du jeu de données issu d'E3N au début de la fenêtre de temps (Q5 en 1997).

Le jeu de données final comprend les informations concernant 90 635 femmes sur les 98 995 qui participent à l'étude E3N. Nous avons en effet choisi d'exclure les femmes qui n'avaient pas répondu au questionnaire qui nous sert de base en 1997 ainsi que les femmes qui n'avaient répondu à aucun questionnaire depuis cette date afin d'être sûr d'avoir un suivi minimal, notamment concernant l'attribut cible.

L'attribut cible « cancer du sein » est considéré comme positif si dans les dix années qui suivent 1997, un cancer du sein est notifié et validé par compte-rendu anatomo-pathologique, négatif sinon. La fenêtre de temps pourra progressivement être allongée au fur et à mesure que les femmes font parvenir les informations sur leur cancer du sein et que celles-ci sont validées. La majorité des cas de cancer du sein sont en effet connus au moment des envois de questionnaires tous les deux à trois ans, un délai maximum auquel il faut ajouter le temps nécessaire à la validation de l'information après contact des médecins.

Attribut cible, le cancer du sein : Nous l'avons décrit dans la partie 1.2.2, page 14, il existe plusieurs types de cancer du sein. Nous avons choisi de considérer uniquement les cancers du sein de type *invasif* et pas les cancers *in situ* qui ne sont considérés que comme des lésions pré-cancéreuses qui n'évoluent pas forcément vers un cancer invasif. Nous avons également vu en partie 1.2.2 qu'il peut exister plusieurs types de cancers invasifs, en fonction de leurs récepteurs hormonaux notamment. Le public cible du score de risque que nous construisons étant constitué de patients et médecins pour qui nous souhaitons construire un score de risque pédagogique et informatif, nous avons choisi de faire un score global pour le cancer du sein et pas un score par type de cancer invasif du sein.

Au total, on compte 2 749 cas de cancer du sein dans le jeu utilisé, soit une classe positive qui compte pour 3,03 % du nombre total d'individus.

Validation croisée : Nous avons choisi, pour les données issues de l'étude E3N, d'utiliser un système de validation croisée en attribuant, tour à tour, 75 % du jeu de données à l'échantillon d'apprentissage et 25 % à l'échantillon de test. Pour y parvenir, le jeu de données est divisé en quatre échantillons. Chacune des femmes du jeu de données est associée aléatoirement à un des quatre échantillons. Ces échantillons constituent tour à tour le jeu de test tandis que les trois autres échantillons constituent le jeu d'apprentissage. Les mesures fournies pour évaluer la performance du score de risque sont obtenues en effectuant la moyenne des quatre mesures obtenues sur les quatre répartitions apprentissage/test. Ces mesures sont associées à un écart-type qui permet de mesurer la dispersion des mesures.

5.2.2.2 Distribution des valeurs

Au sein du jeu de données construit à partir des données E3N, l'âge (arrondi à l'année près) des femmes est réparti selon l'histogramme présenté en figure 5.7. On constate que les femmes les plus jeunes du jeu de données sont plus nombreuses que les femmes les plus âgées. Cette observation peut s'expliquer d'une part par une motivation des femmes à entrer dans l'étude qui dépend de leur âge et, d'autre part, par un biais de recrutement au niveau de la mutuelle qui a servi de source pour le recrutement des participantes.

Les attributs ayant été sélectionnés par rapport à la littérature épidémiologique concernant le cancer du sein et filtrés par un expert, ils peuvent potentiellement tous impacter fortement le risque de cancer du sein ce qui justifie de s'intéresser à la distribution des valeurs pour chacun des attributs que nous utilisons. La figure 5.8, page 130, présente la distribution des valeurs pour les attributs validés.

La discrétisation des valeurs des attributs et les classes générales construites à partir de déclarations plus précises, ont été définies soit de manière à constituer des classes facilement lisibles pour les futurs utilisateurs du score de risque (âge à la ménarche par exemple), soit en fonction de la littérature.

Selon l'épidémiologiste, la distribution des valeurs observées est classique au regard de la population générale et de l'âge moyen des participantes au moment de leur recueil.

5.2.2.3 Espace de recherche

Dans le but de mieux caractériser les données que nous utilisons pour construire le score de risque à destination de la population française, nous avons exploré l'espace de recherche en matière de distance moyenne entre le profil dont on mesure le risque et ses voisins.

Notre estimation du risque de cancer du sein étant basée sur le nombre de personnes malades incluses dans un voisinage, il est intéressant de connaître la distribution des distances auxquelles sont situés les voisins au sein des différents voisinages construits. La figure 5.9 présente la distribution des distances moyennes auxquelles

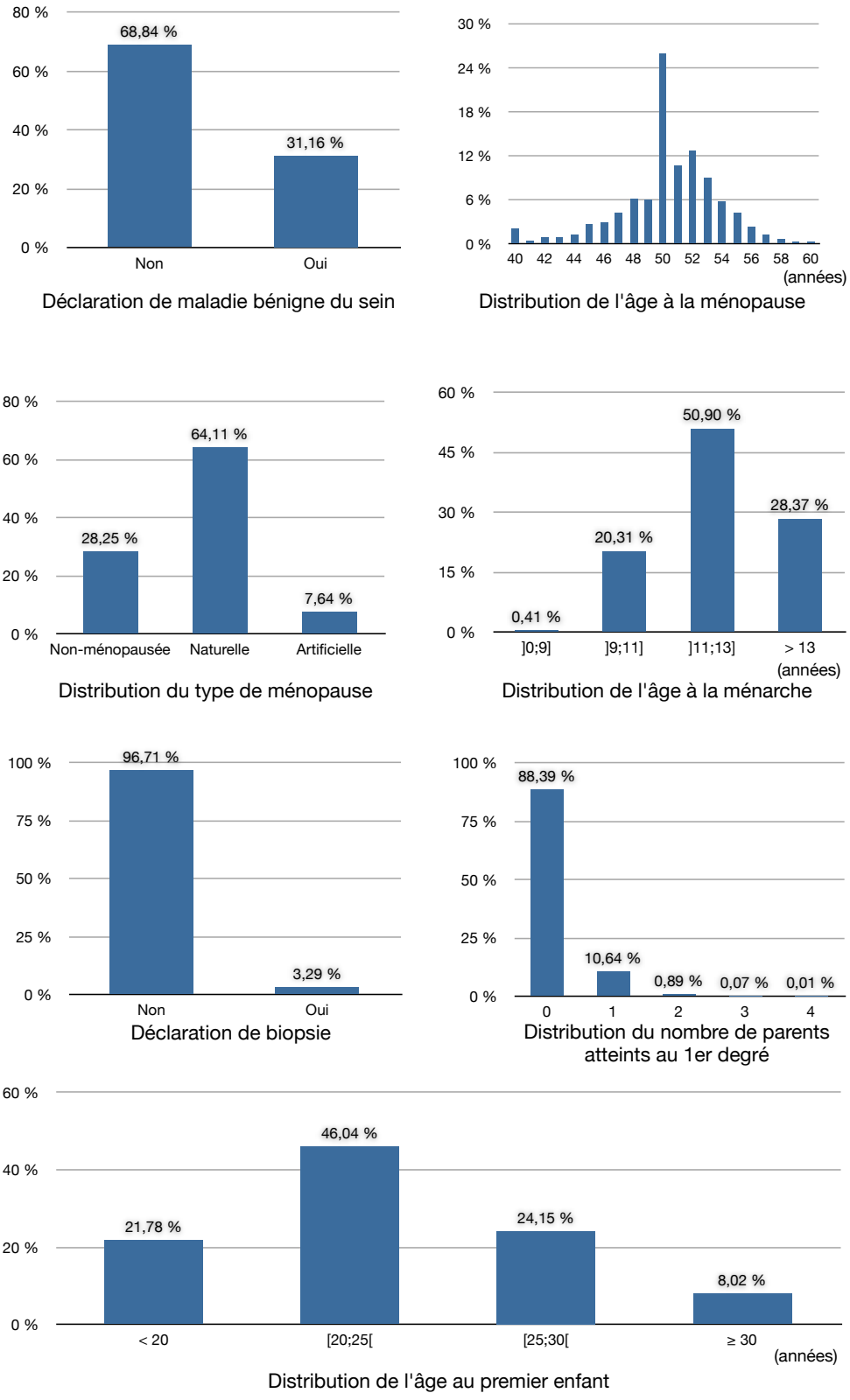

Figure 5.8 : Distribution des valeurs pour les attributs validés.

sont situés les voisins. Les distances retenues ont été calculées sur les combinaisons d'attributs de taille 2 et 3 et seuls les profils pour lesquels au moins un exemple de femme touchée par le cancer du sein sont représentés dans cette distribution moyenne.

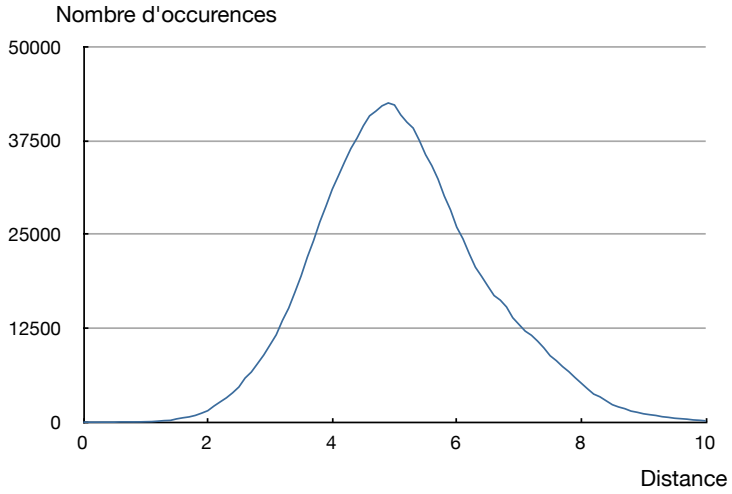

Figure 5.9 : Distribution des distances moyennes au sein des voisinages construits : nombre d'occurrences (en ordonnée) d'un voisin placé à une distance donnée (en abscisse).

On observe que la répartition moyenne des individus dans les voisinages suit une courbe de type loi Normale. On retrouve cette distribution lorsque l'on considère les combinaisons ne comprenant qu'un seul attribut, mais le lissage de la courbe dépend du nombre de modalités disponibles pour décrire l'attribut.

Cette distribution des courbes est un atout pour l'utilisation d'un algorithme des plus proches voisins. En effet, les voisins les plus similaires (ou les plus proches) sont peu nombreux, ainsi l'augmentation de la taille du voisinage considéré permet d'inclure progressivement de plus en plus de voisins. Si le nombre de voisins avait été constant quel que soit la distance euclidienne, nous aurions émis l'hypothèse que la configuration de l'algorithme par l'intermédiaire du nombre de voisins considérés aurait donné lieu à des variations de performances plus marquées au fur et à mesure que la taille du voisinage était augmentée.

Dans cette partie, nous avons expliqué la stratégie de choix des attributs et décrit la distribution de ceux-ci plus l'espace de recherche qu'ils forment, nous avons détaillé le choix de l'attribut cible : cancer du sein invasif non différencié selon son type et nous avons détaillé notre stratégie de validation croisée afin de produire des scores dont les performances, qui permettent de les évaluer, ne soient pas sujettes aux fluctuations statistiques.

5.3 Résultats : production et évaluation des modèles

Nous présentons dans cette partie les performances obtenues d'un point de vue global sur toutes les combinaisons générées, d'abord sur le jeu américain du BCSC qui nous permettra de nous positionner face à un score de risque existant, ensuite sur un jeu issu de l'étude française E3N afin de produire un score qui puisse être utilisé dans le projet de clinique du risque décrit dans le chapitre 3.

5.3.1 Performances pour le jeu du BCSC

Dans le but de pouvoir comparer les performances obtenues par notre utilisation de l'algorithme des proches voisins pour modéliser le risque de cancer du sein, nous avons utilisé le jeu de données mis à disposition du public par le « Breast Cancer Surveillance Consortium » (BCSC) et décrit en partie 5.2.1. En effet, les données de la cohorte E3N n'étant pas publiques et aucun score n'ayant été publié sur ces données, aucune comparaison n'est possible.

Nous avons mesuré la capacité de multiples combinaisons d'attributs à prédire le risque de cancer du sein [Gauthier 11]. Nous avons choisi de tester toutes les combinaisons possibles C_n^t de taille $t = 2$ à $t = 5$ attributs parmi les $n = 10$ attributs retenus par l'épidémiologiste et l'analyste, soit $\sum_{t=2}^{5} C_n^t = \sum_{t=2}^{5} \frac{n!}{t!(n-t)!} = 627$ combinaisons. Chacune de ces combinaisons est testée avec 20 tailles de voisinages différentes, soit 12 540 scores construits et évalués. Nous n'utilisons pas de validation croisée sur ce jeu de données, car [Barlow 06], qui a publié les résultats de référence, a fixé les limites des jeux d'apprentissage et de validation que nous reprenons à l'identique.

Nous présentons d'abord les performances obtenues par taille de combinaison afin de montrer la progression de la capacité à prédire le risque de cancer du sein. Nous montrons ensuite les performances obtenues en classant les combinaisons par AUC décroissante, cet élément de performance étant déterminant dans le choix de la combinaison qui sera fait par l'épidémiologiste et l'analyste. Enfin, nous analysons en détail la combinaison choisie comme score de risque pour le cancer du sein.

5.3.1.1 Performances par taille de combinaisons

Afin de présenter une vue globale des performances obtenues et puisque d'une part, nous observons que ces performances dépendent en partie de la taille de la combinaison d'attributs utilisée et, d'autre part, le nombre d'attributs est un élément important dans la simplicité du score qui sera sélectionné, nous avons choisi de présenter dans le tableau 5.9 les mesures de performances obtenues par tailles de combinaison. Les résultats présentés sont calculés sur le meilleur voisinage de chaque combinaison en termes d'AUC.

En matière d'AUC (probabilité d'attribuer un score plus élevé à un exemple de la classe *malade*), d'ORR (rapport entre la moyenne des scores du premier et dernier

Tableau 5.9 : Évolution de l'AUC en fonction de la taille de la combinaison sur le jeu de données du BCSC.

Taille	Tests	AUC moyenne	AUC médiane	ORR moyen	ORR médian	E/O moyen	E/O médian
2	45	0,567	0,566	2,543	2,187	1,026	1,026
3	120	0,587	0,582	3,330	2,990	1,023	1,025
4	210	0,600	0,598	4,013	3,874	1,018	1,021
5	252	0,608	0,612	4,513	4,504	1,010	1,013

décile) et de rapport E/O moyen (capacité à prédire un nombre de cas de cancer proche du nombre effectivement observé dans le jeu de données), on observe que la performance augmente fortement au fur et à mesure que le nombre d'attributs inclus dans la combinaison augmente jusqu'à quatre attributs. Puis, à partir de quatre attributs, la progression de la performance moyenne ou médiane des indicateurs ralentit. Pour les combinaisons à quatre ou cinq attributs, l'AUC moyen atteint approximativement 0,60. L'ORR atteint un rapport de 4 ce qui signifie qu'il y a 4 fois plus de cas observés de cancer du sein dans le premier décile de score que dans le dernier. La calibration, mesurée par le rapport du nombre estimé d'individus ayant un cancer du sein sur le nombre observé, est excellente puisque même avec un seul attribut, elle atteint 1,026 pour s'améliorer jusqu'à 1,010 au minimum. L'algorithme permet donc de fournir une estimation du nombre de cas de cancer du sein très proche de sa valeur réelle.

Pour compléter cette vision d'ensemble, le tableau 5.10 présente la performance de la meilleure combinaison obtenue, pour chaque taille de combinaison, exprimée sous la forme d'une mesure d'AUC.

Tableau 5.10 : Meilleure combinaison d'attributs par taille de combinaison.

Taille	Meilleure combinaison	AUC	ORR	E/O
2	age+dens	0,635	5,834	1,025
3	age+dens+biop	0,640	6,750	1,025
4	age+dens+biop+precmamm	0,641	6,778	1,016
5	age+dens+biop+precmamm+statmeno	0,640	6,982	1,015

De manière logique, c'est l'âge et la densité qui constituent la meilleure combinaison pour prédire le niveau de risque ce qui confirme la connaissance des experts du domaine (épidémiologiste et analyste) avec un AUC de 0,635. Aux côtés de l'âge, la densité mammaire est connue pour être un facteur de risque important du cancer du sein [McCormack 06, Varghese 12]. La performance maximale en termes d'AUC est atteinte par les combinaisons regroupant l'âge, la densité mammaire et la déclaration de biopsie qui atteint 0,640 sans le résultat de la dernière mammographie et 0,641 avec.

5.3.1.2 Performances par AUC

À l'épidémiologiste et à l'analyste, on présente ces performances globales. Puis on présente les performances, toutes tailles de combinaisons confondues afin de leur permettre de faire un choix : le tableau 5.11 présente les résultats triés par performance d'AUC. Au vu de la difficulté à obtenir l'information relative à la précédente mammographie et vu son faible impact sur la performance des scores qui l'incluent, l'épidémiologiste et l'analyste décident de retirer l'attribut *precmamm* de la liste des scores. Le tableau 5.12 montre la liste filtrée suite à cette décision. Afin d'améliorer l'acceptation du score par les futurs utilisateurs, les experts veulent une combinaison semblable à la combinaison de meilleure performance, mais qui intègre le nombre d'antécédents de cancer du sein. La combinaison qui regroupe les attributs âge de la femme, densité mammaire, nombre biopsies et antécédents familiaux (*age, dens, biop, kdeg1*) possède une performance d'AUC très élevée, excellente en termes de rapport E/O et acceptable en matière d'ORR. Elle a, de plus, l'avantage d'intègrer les attributs habituellement utilisés par les médecins pour évaluer approximativement le niveau de risque de leurs patients : c'est cette dernière qui est choisie avec une AUC maximale obtenue pour 8 000 voisins.

5.3.1.3 Analyse de la combinaison choisie

Notre utilisation de l'algorithme des plus proches voisins avec plusieurs tailles de voisinage permet de choisir la valeur de k qui optimise non seulement la discrimination exprimée sous la forme d'une AUC, mais également exprimée sous la forme de l'ORR et la calibration exprimée sous la forme d'un diagramme de fiabilité et de rapport E/O. Nous présentons ces résultats dans cette partie et les discutons en regard des performances obtenues par [Barlow 06] sur la même base de données.

Statistiques descriptives : Pour une taille de voisinage de 8 000 individus, la combinaison âge, densité mammaire, la déclaration de biopsie et nombre d'antécédents de cancer du sein a permis de calculer des scores basés sur des voisinages intégrant en moyenne 58 femmes atteintes par un cancer du sein. La valeur médiane est de 53 femmes atteintes. Les voisinages construits pour calculer les prévalences qui constituent les scores de risque ont intégré 5 femmes malades au minimum et 234 au maximum.

Mesures de discrimination : La figure 5.10 présente l'évolution de l'AUC pour cette combinaison sur une large gamme de valeurs de taille de voisinages. On constate que l'AUC maximale n'est pas atteinte pour une seule taille de voisinage, mais sur une gamme de valeurs de taille de voisinages : l'AUC dépasse 0,638 pour les tailles de voisinage comprises entre 7 500 et 8 500 voisins et 0,637 pour les tailles de voisinage comprises entre 6 500 et 9 500 voisins environ. Les performances chutent ensuite quand on augmente la taille du voisinage, en effet, plus la taille du

Tableau 5.11 : Combinaisons triées par meilleure performance d'AUC avant filtrage.

Combinaisons avant filtrage	AUC	E/O
age, dens, biop, precmamm	0,641	1,016
age, dens, biop	0,640	1,025
age, dens, biop, agemeno, precmamm	0,640	1,015
age, dens, biop, agemeno	0,640	1,025
age, dens, biop, kdeg1	0,638	0,993
age, dens, biop, agemeno, ths	0,637	1,026
age, dens, biop, kdeg1, ths	0,637	0,997
age, dens, biop, ths	0,637	1,024
age, dens, biop, precmamm, ths	0,637	1,010
age, dens, biop, kdeg1, precmamm	0,636	0,978
age, dens, biop, agemeno, typemeno	0,636	1,026
age, dens, biop, agenai	0,636	1,011
age, dens, biop, agemeno, kdeg1	0,636	0,990
age, dens, biop, typemeno	0,636	1,027
age, dens, biop, agenai, imc	0,635	0,993

Tableau 5.12 : Combinaisons triées par meilleure performance d'AUC après filtrage.

Combinaisons après filtrage	AUC	E/O
age, dens, biop	0,640	1,025
age, dens, biop, agemeno	0,639	1,025
age, dens, biop, kdeg1	0,638	0,993
age, dens, biop, agemeno, ths	0,637	1,026
age, dens, biop, kdeg1, ths	0,637	0,997
age, dens, biop, ths	0,637	1,024
age, dens, biop, agemeno, typemeno	0,636	1,026
age, dens, biop, agenai	0,636	1,011
age, dens, biop, agemeno, kdeg1	0,636	0,990
age, dens, biop, typemeno	0,636	1,027
age, dens, biop, agenai, imc	0,635	0,993
age, dens, biop, imc	0,635	1,015
age, dens, typemeno	0,635	1,024
age, dens	0,635	1,025
age, dens, agemeno, typemeno	0,635	1,024

Tableau 5.13 : Correspondance entre nom abrégé et nom complet de l'attribut.

Nom abrégé	Attribut	Nom abrégé	Attribut
age	Âge de la femme	agemeno	Âge de la femme à la ménopause
agenai	Âge à la naissance du 1er enfant	typemeno	Type de ménopause
imc	indice de masse corporelle	dens	Densité du sein
biop	Déclaration de biopsie	kdeg1	Nombre d'antécédents au premier degré
ths	Traitement hormonal substitutif		

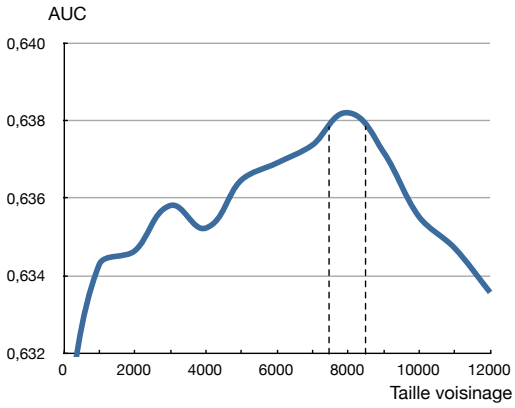

Figure 5.10 : Évolution de l'AUC en fonction de la taille du voisinage.

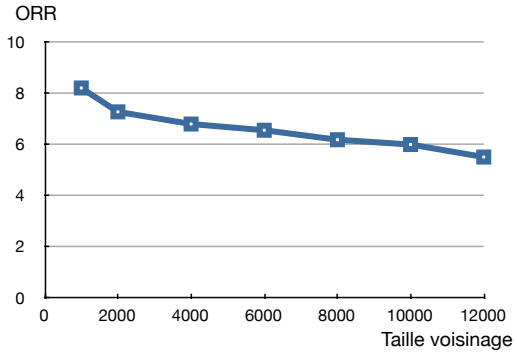

Figure 5.11 : Évolution de l'ORR en fonction de la taille du voisinage.

voisinage augmente, plus des voisins différents de la personne dont on veut mesurer le risque sont ajoutés au voisinage, diminuant mécaniquement les performances de la combinaison testée.

Pour la combinaison choisie, avec une taille de voisinage de 8 000 voisins, l'ORR calculé est de 6,176. Pour la combinaison choisie, il y a donc, lorsque les individus du jeu de validation sont classés en fonction du score qui leur est attribué, six fois plus de malades du cancer du sein dans le premier décile de score que dans le dernier. L'évolution de l'ORR est présentée sur la figure 5.11, page 136, sur le même intervalle que l'AUC de la figure 5.11. Au contraire de l'AUC, les meilleurs rapports E/O entre déciles sont obtenus pour les tailles de voisinage les plus petites, puis au fur et à mesure de l'augmentation de la taille des voisinages, l'ORR diminue.

Mesures de calibration : Le rapport du nombre estimé d'individus malades par rapport au nombre observé est de 0,993 pour la combinaison et le nombre de voisins choisis. Ainsi, notre utilisation de l'algorithme des plus proches voisins prédit 2 272 femmes malades dans le jeu de validation tandis qu'en réalité, celui-ci en regroupe 2 286. La calibration est donc excellente, ce qui est confirmé pour le diagramme de fiabilité qui précise le rapport E/O par décile de risque (voir figure 5.12). Théoriquement, le meilleur diagramme de fiabilité possible est une diagonale et on observe que la courbe tracée à partir de la comparaison des nombres de cancers du sein prédits et observés est proche de cette diagonale symbolisant la calibration parfaite.

À partir de ces données, l'algorithme des plus proches voisins permet de produire un score très bien calibré. Nous émettons l'hypothèse que c'est la nature même de l'algorithme qui permet d'obtenir une telle précision dans la prédiction du nombre de cas de cancer du sein. En effet, le score attribué aux profils consiste à calculer le rapport du nombre de personnes atteintes sur le nombre de personnes du groupe

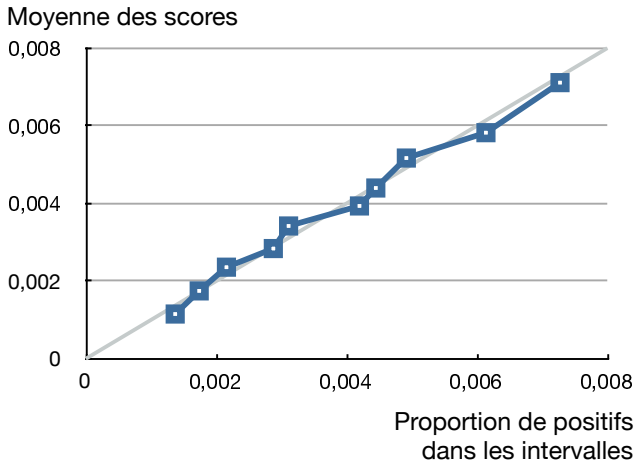

Figure 5.12 : Diagramme de fiabilité de la combinaison choisie pour un voisinage de 8 500 femmes.

considéré. Or la calibration étant précisément la mesure de la capacité à prédire ce nombre de personnes atteintes, cette dernière est en conséquence excellente.

Discussion des résultats : Sur le même jeu de données et comme expliqué en partie 2.2.1.2, page 36, [Barlow 06] a choisi de créer deux scores de risque différents en fonction du statut ménopausique de la femme. Pour la prédiction des cancers invasifs, le modèle pour les femmes en préménopause inclut quatre attributs pour une performance d'AUC de 0,633 et un rapport E/O de 0,91 tandis que le modèle pour les femmes en postménopause inclut 10 attributs pour une performance de 0,628 d'AUC et un rapport E/O que nous estimons à 1,01 en fonction des informations partielles fournies dans l'article.

Avec seulement quatre attributs dans un seul score, l'algorithme des plus proches voisins que nous utilisons permet de proposer un score de risque pour le cancer du sein qui affiche une performance comparable (0,638 contre 0,628 et 0,633) toutes femmes confondues en matière de discrimination et une meilleure performance en matière de calibration avec un rapport E/O de 0,99 contre 0,91 et 1,01 pour les scores proposés par Barlow *et al.*

En outre, le processus que nous proposons permet d'adapter la composition du score en termes d'attributs au contexte d'utilisation du score, ce qui conduit à ne pas choisir le score de meilleure performance (0,641 quand aucun attribut n'est éliminé dans le processus) pour adapter la composition du score sans dégrader trop fortement les performances affichées.

Notre objectif est maintenant de tester notre processus associé à l'algorithme des plus proches voisins sur une autre base de données afin de confirmer ces premiers résultats et de construire un score de risque pour la population française.

5.3.2 Performances pour le jeu issu de l'étude E3N

Après avoir mesuré les performances obtenues avec l'algorithme des proches voisins sur une base publique dans le but de comparer les performances avec celles de la littérature, nous utilisons un jeu de données issu de l'étude française E3N (voir page 125) pour proposer à des femmes françaises d'utiliser un tel score de risque, par exemple dans le cadre de la clinique de risque décrite en partie 3.1.1.1, page 60.

Toujours en utilisant le système informatique décrit en partie 4.4, page 103, nous avons mesuré la capacité de multiples combinaisons d'attributs à prédire le risque de cancer du sein[Gauthier 12]. Cette fois, nous avons choisi de tester toutes les combinaisons possibles C_n^t de taille $t = 2$ à $t = 5$ attributs parmi les $n = 8$ attributs retenus par l'expert du domaine, soit $\sum_{t=2}^{5} C_n^t = \sum_{t=2}^{5} \frac{n!}{t!(n-t)!} = 210$ combinaisons. Chacune de ces combinaisons est testée avec 15 tailles de voisinages uniformément réparties de 1 000 à 15 000 voisins, soit au total 3 150 scores construits et évalués sur chacun des quatre jeux de validation croisée (voir partie 5.2.2.1, page 129). Le nombre d'attributs retenus par l'épidémiologiste et l'analyste est plus faible que pour le jeu de données précédent, car ceux-ci sont des experts du domaine qui ont déjà une connaissance particulière de l'étude E3N d'où proviennent les données et une idée précise des attributs qu'ils souhaitent intégrer dans la combinaison qu'ils veulent utiliser comme score de risque.

De nouveau, nous présentons les performances obtenues par taille de combinaison, puis les performances obtenues en classant les combinaisons par AUC décroissante et la combinaison choisie.

5.3.2.1 Performances par taille de combinaisons

Afin de présenter une vue globale des performances obtenues, nous montrons dans le tableau 5.14 les résultats regroupés par taille de combinaison testée. Les résultats présentés sont des moyennes des résultats par taille en utilisant pour chaque combinaison, la taille de voisinage qui conduit à la meilleure AUC. Comme pour le jeu américain BCSC, les performances moyennes et médianes d'AUC augmentent nettement pour les combinaisons à deux ou trois attributs, avant de se stabiliser autour d'une AUC de 0,60. De même, la hausse de l'ORR moyen et médian est forte pour les petites tailles de combinaisons avant de ralentir à partir de quatre attributs.

La calibration moyenne connaît un plafond avec 0,985 de moyenne pour les combinaisons à deux attributs, elle diminue ensuite pour les combinaisons à cinq attributs. La performance est tout de même excellente puisque l'algorithme des proches voisins permet de prédire entre 660 et 688 cas de cancer du sein pour 688 observés.

Le tableau 5.15 présente la combinaison de meilleure AUC par taille de combinaison. De manière logique, parmi les attributs testés, on trouve l'âge qui domine le classement des combinaisons avec la déclaration de maladie bénigne du sein (*mbs*). L'âge est connu pour un être un bon prédicteur du risque de cancer du sein. Le fait

Tableau 5.14 : Évolution de l'AUC en fonction de la taille de la combinaison sur le jeu de données issu de l'étude E3N.

Taille	Tests	AUC moyenne	AUC médiane	ORR moyen	ORR médian	E/O moyen	E/O médian
2	28	0,561	0,555	1,995	2,005	0,985	1,000
3	56	0,578	0,577	2,407	2,342	0,979	0,991
4	70	0,591	0,593	2,914	2,958	0,976	0,984
5	56	0,602	0,605	3,182	3,015	0,962	0,964

d'avoir subi une maladie bénigne également, malgré les questions qui se posent sur le continuum lésionnel évoqué page 16.

C'est ensuite l'âge à la ménopause (*agemeno*) qui permet d'augmenter le plus les performances, ce qui est cohérent avec la littérature et l'impact important de l'imprégnation hormonale sur le risque de cancer du sein, voir page 19.

Le nombre d'antécédents au premier degré apporte ensuite assez d'information pour augmenter la discrimination du score à 0,625 en termes d'AUC, ce qui cohérent avec la littérature qui rapporte un risque plus élevé d'un facteur 3,9 pour une personne ayant trois antécédents de cancer du sein au premier degré relativement aux personnes n'en ayant aucun [Lichtenstein 00].

C'est enfin le type de ménopause qui permet d'augmenter légèrement les performances. En effet, en cas de ménopause artificielle due à l'ablation des ovaires, l'imprégnation hormonale est fortement diminuée, ce qui explique le léger gain en discrimination de cet attribut.

5.3.2.2 Performances par AUC

Après avoir présenté ces performances globales à l'épidémiologiste et à l'analyste afin de vérifier la cohérence des résultats, une liste des meilleures combinaisons, sans condition sur la taille de celles-ci, leur est présentée afin d'initier le processus de choix de la combinaison.

Tableau 5.15 : Meilleure combinaison d'attributs par taille de combinaison pour le jeu de données issu de l'étude E3N.

Taille	Meilleure combinaison	AUC	ORR	E/O
2	age+mbs	0,604	3,080	1,004
3	age+mbs+agemeno	0,618	3,999	0,992
4	age+mbs+agemeno+kdeg1	0,625	4,769	0,993
5	age+mbs+agemeno+kdeg1+typemeno	0,628	4,645	0,976

Le tableau 5.16 présente les meilleures combinaisons classées par AUC sans distinction de nombre d'attributs utilisé. Le tableau 5.18 rappelle la signification des abréviations utilisées. Parmi toutes les combinaisons testées, la combinaison qui possède la meilleure AUC est constituée des attributs âge (*age*), maladies bénignes du sein (*mbs*), âge à la ménopause (*agemeno*), antécédents au premier degré (*kdeg1*) et type de ménopause (*typemeno*) pour les femmes ménopausées avec une AUC de 0,628 et un rapport E/O de 0,990. Les experts éliminent les combinaisons qui n'utilisent pas l'âge pour calculer le risque. Malgré les performances correctes de ces combinaisons, il serait difficile de faire accepter aux utilisateurs un score de risque qui ne comprenne pas l'âge de la femme.

Une nouvelle liste est donc générée et présentée dans le tableau 5.17. Les combinaisons en tête du classement par AUC sont considérées comme des possibilités qui répondent aux objectifs de performance et de lisibilité. La troisième combinaison de la liste est privilégiée pour plusieurs raisons. Son AUC est élevée et comparable aux deux autres (la différence est de seulement 0,003), mais elle a l'avantage d'être calculable à partir de quatre attributs seulement, un critère de simplicité important. Enfin, même si la majorité des mesures de calibration sont excellentes, la calibration mesurée pour cette troisième combinaison est meilleure que celle des deux précédentes avec un rapport E/O calculé à 0,993 contre 0,976 et 0,990 pour les précédentes. Le coût d'obtention des informations n'est pas un critère de choix pertinent pour la sélection d'une combinaison, car le contexte a été pris en compte lors de l'étape du choix des attributs.

Ce type de combinaison conviendrait pour une utilisation auprès du grand public comme dans le projet d'espace pédagogique du Centre Hygée évoqué page 60. Malgré une calibration en léger retrait, la deuxième combinaison pourrait également convenir à un projet de type clinique du risque évoqué page 60 en raison de la présence de l'attribut indiquant si une femme a subi précédemment une biopsie (*biop*), un marqueur de risque du type de la maladie bénigne du sein qui vient renforcer ce dernier, toutes les femmes déclarant une ou plusieurs maladies bénignes du sein n'ayant pas forcément subi des biopsies.

Dans la partie suivante, nous détaillons donc la combinaison âge, maladies bénignes du sein, âge à la ménopause et nombre d'antécédents familiaux de cancer du sein au premier degré qui est suggérée en sortie de processus, qui affiche une performance de discrimination de 0,625 et une calibration de 0,993 mesurée par le rapport entre le nombre de cas de cancer estimé et le nombre observé de cas de cancer.

5.3.2.3 Analyse de la combinaison suggérée

Pour analyser cette combinaison plus en détail, nous fournissons quelques statistiques descriptives afin de mieux cerner le fonctionnement de l'algorithme sur la population. Nous analysons la discrimination, en termes d'AUC et d'ORR, et la calibration en termes de rapport E/O que nous détaillons par un diagramme de fiabilité. Nous discutons ensuite ces résultats en les comparant aux autres résultats

Tableau 5.16 : Combinaisons triées par meilleure performance d'AUC avant filtrage.

Combinaisons avant filtrage	AUC	E/O
age, mbs, agemeno, kdeg1, typemeno	0,628	0,976
age, mbs, agemeno, kdeg1, biop	0,628	0,990
age, mbs, agemeno, kdeg1	0,625	0,993
age, mbs, agemeno, typemeno, biop	0,624	0,989
age, mbs, agemeno, kdeg1, agenai	0,623	0,974
mbs, agemeno, kdeg1, typemeno, biop	0,622	0,959
age, mbs, agemeno, kdeg1, menarche	0,621	0,987
age, mbs, agemeno, biop	0,621	0,998
age, mbs, agemeno, typemeno	0,621	0,990
age, mbs, kdeg1, typemeno, biop	0,620	0,932
age, mbs, agemeno, biop, agenai	0,620	0,983
age, mbs, kdeg1, biop, agenai	0,619	0,946
mbs, agemeno, kdeg1, typemeno	0,619	0,968
age, mbs, agemeno	0,618	0,992
age, mbs, kdeg1, biop, menarche	0,617	0,896

Tableau 5.17 : Combinaisons triées par meilleure performance d'AUC après filtrage.

Combinaisons après filtrage	AUC	E/O
age, mbs, agemeno, kdeg1, typemeno	0,628	0,976
age, mbs, agemeno, kdeg1, biop	0,628	0,990
age, mbs, agemeno, kdeg1	0,625	0,993
age, mbs, agemeno, typemeno, biop	0,624	0,989
age, mbs, agemeno, kdeg1, agenai	0,623	0,974
age, mbs, agemeno, kdeg1, menarche	0,621	0,987
age, mbs, agemeno, biop	0,621	0,998
age, mbs, agemeno, typemeno	0,621	0,990
age, mbs, kdeg1, typemeno, biop	0,620	0,932
age, mbs, agemeno, biop, agenai	0,620	0,983
age, mbs, kdeg1, biop, agenai	0,619	0,946
age, mbs, agemeno	0,618	0,992
age, mbs, kdeg1, biop, menarche	0,617	0,896
age, mbs, agemeno, typemeno, agenai	0,617	0,962
age, mbs, kdeg1, typemeno	0,617	0,975

Tableau 5.18 : Correspondance entre nom abrégé et nom complet de l'attribut.

Nom abrégé	Attribut	Nom abrégé	Attribut
age	Âge de la femme	agemeno	Âge de la femme à la ménopause
agenai	Âge à la naissance du 1er enfant	typemeno	Type de ménopause
menarche	Âge de la femme à la ménarche	mbs	Déclaration de maladie bénigne du sein
biop	Déclaration de biopsie	kdeg1	Nombre d'antécédents au premier degré

de la littérature.

Statistiques descriptives : La combinaison suggérée est constituée des attributs âge, maladie bénigne du sein, âge à la ménopause et nombre d'antécédents familiaux de cancer du sein au premier degré. La taille de voisinage utilisée pour calculer les prévalences qui servent de score est de 2 000 voisins. En moyenne sur tous les ensembles d'apprentissage, 87 femmes malades sont incluses dans les voisinages constitués à partir des calculs de distances tandis que la médiane est de 88 femmes malades (22 au minimum et 154 au maximum). L'ordre de grandeur de ces chiffres est cohérent avec ceux obtenus sur la base du BCSC avec 58 femmes atteintes par le cancer du sein en moyenne par voisinage.

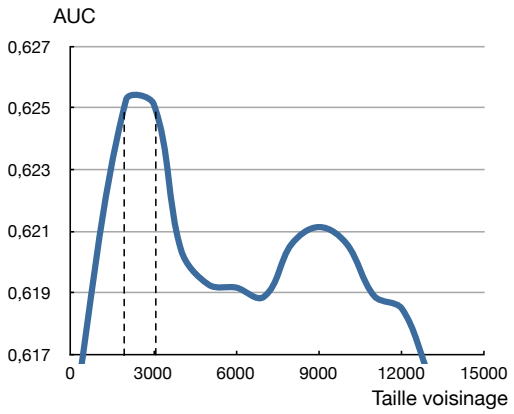

Figure 5.13 : Évolution de l'AUC de la combinaison *age, mbs, agemeno, kdeg1* en fonction de la taille du voisinage.

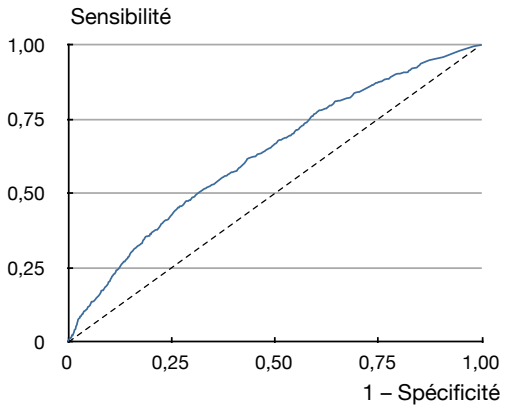

Figure 5.14 : Courbe ROC de la combinaison *age, mbs, agemeno, kdeg1* pour un voisinage de 2 000 femmes.

Mesures de discrimination : La figure 5.13 présente l'évolution de l'AUC pour la combinaison suggérée sur une large gamme de valeurs de taille de voisinage. Comme pour le jeu précédent, on constate que l'AUC maximale n'est pas atteinte pour une taille précise de voisinage en dehors de laquelle les performances seraient largement inférieures, mais sur une gamme de valeurs de taille de voisinage. Atteindre une valeur d'AUC sur une gamme de tailles de voisinage permet d'assurer que la performance affichée pour une combinaison n'est pas le résultat d'une fluctuation statistique. L'AUC dépasse 0,625 pour les tailles de voisinage comprises entre 2 000 et 3 000 voisins (voir la courbe ROC correspondante figure 5.14), soit environ 4,5 % de la taille du jeu de validation. Les performances chutent ensuite quand on augmente la taille du voisinage, car des voisins situés à des distances de plus en plus grandes. Ces voisins ajoutés sont ainsi de plus en plus différents ce qui entraîne une chute des performances mis à part un léger rebond observé autour de 9 000 voisins pour lequel nous proposons une explication ci-dessous.

Cette valeur d'AUC a été calculée en utilisant une validation croisée. Comme détaillé page 129, quatre valeurs d'AUC ont donc été calculées : l'écart type mesuré sur ces quatres valeurs est de 0,005 ce qui dénote une faible dispersion des valeurs entre les mesures. L'observation de l'évolution de cette valeur d'écart type en fonction de la taille du voisinage montre une augmentation de la dispersion des valeurs autour de 10 000 voisins qui correspond à l'augmentation d'AUC observée. Nous émettons l'hypothèse que la remontée des performances est due à une fluctuation statistique d'échantillonnage due au jeu de données et à la répartition des valeurs dans les jeux d'apprentissage et de validation.

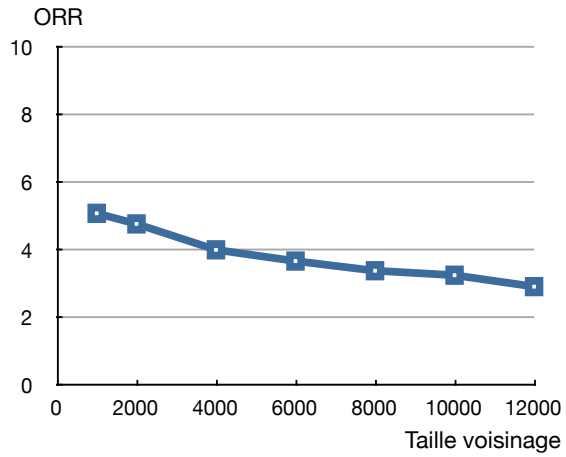

Figure 5.15 : Évolution de l'ORR de la combinaison *age, mbs, agemeno, kdeg1* en fonction de la taille du voisinage.

Pour la combinaison retenue avec 2 000 voisins, l'ORR observée est de 4,769. Pour cette combinaison, il y a donc, lorsque les individus du jeu de validation sont classés en fonction du score qui leur est attribué, près de 4,8 fois plus de malades du cancer du sein dans le premier décile de score que dans le dernier. L'évolution de l'ORR est représentée sur la figure 5.15 sur le même intervalle de taille de voisinage que l'AUC de la figure 5.15. On constate le même phénomène que pour le jeu BCSC précédemment étudié, au contraire de l'AUC, les meilleurs rapports entre déciles sont obtenus pour les tailles de voisinage les plus petites, puis au fur et à mesure de l'augmentation de la taille des voisinages, l'ORR diminue.

La dispersion statistique des mesures d'ORR est constante en fonction des tailles de voisinages observées.

Mesures de calibration : Le rapport du nombre estimé d'individus malades par rapport au nombre observé est mesuré à 0,993 pour 2 000 voisins. Ainsi, pour le point haut de la courbe d'AUC, notre utilisation de l'algorithme des plus proches voisins prédit environ 681 femmes malades dans le jeu de validation tandis qu'en réalité,

celui-ci en regroupe en moyenne 687 sur les jeux de validation croisée. Comme pour le jeu BCSC, la calibration est excellente, ce qui est confirmé pour le diagramme de fiabilité qui précise le rapport E/O par décile de risque (voir figure 5.16). Théoriquement, le meilleur diagramme de fiabilité possible est une diagonale et on observe que la courbe tracée à partir de la comparaison des nombres de cancers du sein prédits et observés est très proche de cette diagonale symbolisant une excellente calibration. Ce score est donc très bien calibré.

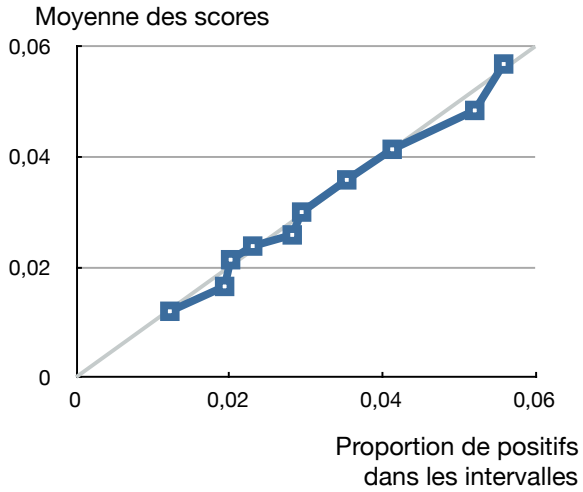

Figure 5.16 : Diagramme de fiabilité de la combinaison *age, mbs, agemeno, kdeg1* pour un voisinage de 2 000 femmes.

Discussion des résultats : Les performances relevées sur le précédent jeu de données (jusqu'à 0,641 d'AUC sur le jeu BCSC) sont légèrement meilleures que celles mesurées sur les données issues de l'étude E3N (jusqu'à 0,628), y compris si le processus de choix de la composition de la combinaison est mis de côté. Nous expliquons cette différence de performance par les attributs à notre disposition pour estimer le niveau de risque. En effet, la densité mammaire, non disponible à l'heure actuelle dans l'étude E3N, est un facteur de risque important dans le cancer du sein. Posséder cette information, même si elle est obtenue avec un système de mesure perfectible comme c'est le cas pour les données du jeu de données BCSC, permettrait d'augmenter les performances.

En revanche si l'on compare les performances des modèles construits grâce aux données de l'étude E3N aux modèles de la littérature utilisant des attributs semblables, les performances obtenues sont comparables. Par exemple, pour le modèle de [Gail 89] dans sa version 2 par [Costantino 99], le modèle le plus utilisé pour estimer le risque de cancer du sein (voir page 33), [Rockhill 01] a mesuré une AUC de 0,58 pour une combinaison incluant l'âge, la déclaration de biopsie, les antécédents

familiaux, l'âge à la ménarche* et l'âge au premier enfant quand la même combinaison a une performance de 0,587 en matière d'AUC avec l'algorithme des plus proches voisins. Une mesure de discrimination très proche de celle que nous observons. En ajoutant la densité dans le modèle de Gail, [Chen 06] modifie fortement l'équation de régression et la mesure d'AUC augmente à 0,643, soit une valeur très proche de celles obtenues par [Barlow 06] et nous même lorsque la densité mammaire est utilisée.

La calibration est toutefois largement améliorée et nous émettons l'hypothèse que c'est grâce à l'utilisation de l'algorithme des plus proches voisins. En effet, en fonction du modèle de Gail utilisé et des sous-populations utilisées pour tester le score, la calibration varie de 0,84 à 1,03 pour le rapport E/O alors que le modèle basé sur les plus proches voisins permet d'atteindre 0,993 pour la combinaison choisie.

Ces résultats suggèrent que l'amélioration des performances passe par l'utilisation de nouveaux attributs, facteurs ou marqueurs de risque, qui permettent de mieux discriminer les femmes à haut risque de cancer du sein sans dégrader la calibration. C'est dans cette optique que les responsables de l'étude E3N ont lancé en 2009 un projet de recueil rétrospectif de la densité mammaire chez certaines femmes de la cohorte pour construire une étude cas-témoin*. Environ 3 500 femmes victimes d'un cancer du sein, plus un à trois témoins par cas de cancer du sein, ont été invitées à retourner les clichés mammographiques en leur possession. Une logistique importante a été mise en place pour permettre la numérisation des éléments envoyés. Parallèlement, différentes méthodes de calcul de la densité sont testées afin d'obtenir des valeurs d'attributs fiables qui puissent être utilisables dans des modèles de risque.

Afin de compléter notre analyse de cette suggestion de combinaison pour le score de risque, nous avons comparé la manière dont les individus sont classés par score en fonction de la méthode de modélisation utilisée.

5.3.3 Comparaison avec une régression logistique

Si nous avons pu confronter nos résultats à ceux obtenus sur le jeu américain du BCSC par [Barlow 06], nous n'avons pas de référence publiée pour le score sur le jeu de données issu de l'étude E3N. Nous avons donc choisi de modéliser le risque de cancer du sein sur le même jeu de données issu de l'étude E3N avec une régression logistique non conditionnelle. Cette méthode est utilisée par [Gail 89] dans son modèle le plus utilisé, auquel il ajoute une correction mineure pour prendre en compte les risques compétitifs comme le décès par d'autres causes que le cancer du sein. Nous ne l'avons pas reprise dans les résultats présentés ci-dessous.

Sur les quatre jeux d'apprentissage utilisés jusqu'à présent pour le jeu de données issu de l'étude E3N, nous avons déterminé les équations d'une régression logistique dont nous nous sommes servis pour attribuer un score aux femmes des quatre jeux de validation. Pour la combinaison suggérée précédemment (âge, maladie bénigne, âge à la ménopause et nombre d'antécédents), l'AUC mesurée est de 0,614 avec la

régression logistique (écart type de 0,003 pour les quatre mesures) contre 0,625 avec notre utilisation de l'algorithme des plus proches voisins.

Nous réalisons ensuite la moyenne des quatre scores calculés pour les deux méthodes, nous assignons les deux scores aux 22 661 femmes d'un ensemble de validation choisi parmi les quatre et nous classons les femmes par ordre décroissant de score. Nous mesurons la corrélation entre le classement issu de la régression logistique et le classement issu de l'utilisation de l'algorithme des plus proches voisins. Nous divisons ensuite la population en quartiles et nous mesurons la concordance entre les classements effectués.

Corrélation entre les classements : Le coefficient de corrélation de Pearson ρ, calculé entre les deux classements est de 0,86. Comme indiqué en partie 2.3.4.1, page 55, si ρ vaut 0, il n'existe pas de corrélation et si ρ vaut 1, la corrélation est parfaite.

La corrélation est donc forte entre les deux classements effectués suite à l'utilisation des deux méthodes d'estimation. Pour voir une vision plus fine de l'évolution du classement des femmes en fonction de leur niveau de risque, nous complétons cette mesure par l'utilisation de l'indice de concordance κ.

Concordance entre les classements : Le tableau 5.19 présente les réaffectations subies par les femmes par quartile en fonction de la méthode de modélisation utilisée.

Un tel tableau dans lequel seules les diagonales grisées comporteraient des effectifs non-nuls, montrerait que, quel que soit le score utilisé, les femmes n'auraient pas subi de réaffectation de quartile. On constate que ce n'est pas le cas et on utilise donc l'indice de concordance κ pour évaluer le taux de concordance, c'est-à-dire le taux de reclassification d'un score à l'autre.

Tableau 5.19 : Matrice de concordance pour 22 661 femmes classées par quartiles de valeurs en fonction des scores obtenus par régression logistique et algorithme des plus proches voisins.

		Proches voisins				
		quartile 1	quartile 2	quartile 3	quartile 4	Total
Régression	quartile 1	4 251 (18,76 %)	796 (3,51 %)	618 (2,73 %)	0	5 665
	quartile 2	944 (4,17 %)	2 745 (12,11 %)	1976 (8,72 %)	0	5 665
	quartile 3	470 (2,07 %)	2117 (9,34 %)	2 275 (10,04 %)	804 (3,55 %)	5 666
	quartile 4	0	7 (0,03 %)	797 (3,52 %)	4 861 (85,81 %)	5 665
	Total	5 565	5 565	5 666	5 565	22 661

L'indice κ mesuré avec pondération est de 0,66, ce qui correspond à une bonne concordance d'après l'échelle de Landis et Koch (voir page 57). Le détail du calcul

de l'indice de concordance κ est disponible en partie 2.3.4.2, page 56.

Vu la mesure de concordance et le taux de corrélation, nous considérons que la similitude entre les classements effectués par les deux méthodes utilisées est avérée. Un algorithme du type des plus proches voisins permet donc d'assigner les scores de manière cohérente avec la régression logistique, une méthode populaire dans le domaine de la création de scores de risque pour le cancer du sein. Il est en revanche impossible de conclure quant à la supériorité d'un classement sur l'autre avec ce type de mesure.

Au cours de ce chapitre, nous avons détaillé les jeux de données utilisés et la méthode d'estimation du risque choisie, afin d'expérimenter la construction d'un score de risque grâce au processus proposé et à l'implantation de l'architecture qui le supporte. Nous avons proposé des scores de risque dont la performance de discrimination est comparable, voire supérieure à celle des scores proposés dans la littérature et dont la performance de calibration est manifestement supérieure. L'utilisation du processus proposé au chapitre 3 a, en outre, permis de prendre en compte le contexte d'utilisation des scores et la disponibilité des attributs pour choisir la nature et le nombre d'attributs qui composent ces scores sans dégrader trop fortement les performances. Les scores proposés pour le cancer du sein sont donc globalement plus performants tout en restant compréhensibles, car ils sont calculés avec une méthode des plus proches voisins facilement explicable aux utilisateurs et comportant moins d'attributs que les scores présentés dans la littérature.

Conclusion

Si la lutte contre les grandes maladies comme les cancers, les maladies cardio-vasculaires ou le déclin cognitif passe par l'amélioration et la création de traitements médicaux, la prévention est également un moyen de diminuer l'impact et l'incidence de ces grandes maladies. En particulier, une prévention ciblée sur les profils de personnes les plus à risque permet de diminuer la mortalité due à ces maladies, d'une part en communiquant sur les bons comportements pour éviter l'apparition de la maladie et, d'autre part, en intensifiant le dépistage des maladies pour les diagnostiquer plus tôt afin d'augmenter les chances de réussite des traitements.

L'utilisation de scores de risque participe à ces deux axes importants de la prévention. En effet, l'utilisation de scores de risque permet de mettre en évidence de mauvais comportements lorsque les facteurs de risque d'une maladie sont modifiables (meilleure alimentation ou augmentation de l'activité physique par exemple) et de proposer des manières de corriger ces mauvais comportements. L'utilisation des scores de risque permet également de détecter les profils à risque et d'adapter le dépistage des maladies pour ces personnes (modulation de la date de début du suivi mammographique des femmes pour le cancer du sein par exemple).

Principaux résultats

Nous avons proposé un processus de création de scores de risque dans le domaine de la santé qui permet de corriger les faiblesses des scores de risque existants en ce qui concerne son adaptation au contexte tout en proposant des performances équivalentes ou supérieures. Nous avons conçu une architecture pour supporter ce processus et nous nous en sommes servis pour expérimenter la construction de scores de risque sur deux bases de données différentes, une base américaine mise à disposition sur internet et une base française gérée par l'INSERM.

Après avoir fait le constat que les scores de risque dans le domaine de la santé sont peu flexibles et difficilement adaptables au contexte d'utilisation, nous proposons un processus de création de scores de risque basé sur un modèle de processus de fouille de données reconnu. Le processus que nous proposons permet de prendre en compte, dès le début de la conception du score de risque, les contraintes liées à son contexte d'utilisation, notamment en fonction de la maladie visée, du contexte général d'utilisation et du type d'utilisateur. Afin de montrer la viabilité de cette proposition de processus de création de scores de risque, nous l'avons appliquée dans le domaine du cancer du sein.

Dans le but de réaliser des expérimentations qui permettent de tester ce processus, nous avons proposé une architecture de système d'information qui supporte

les processus de production et d'utilisation des scores de risque proposés. Pour cela, nous avons précisément décrit les processus mis en jeux et avons choisi d'aligner la vue fonctionnelle sur la vue métier. En alignant les cas d'utilisation sur les activités et la définition des entités participantes sur les données métiers, nous avons pu, à l'aide de la conception des îlots fonctionnels, déduire les données fonctionnelles, les diagrammes de séquence puis les relations entre îlots et entre données fonctionnelles. Nous avons utilisé ces derniers pour concevoir des composants applicatifs qui permettent de développer un système dont le but est de faciliter l'expérimentation du processus en conditions réelles.

En suivant le processus proposé et en utilisant les composants applicatifs développés pour mettre en œuvre une partie du système d'information proposé, nous avons mené des expériences pour mesurer la performance d'un algorithme de prédiction du risque de cancer du sein.

Nous avons ainsi montré qu'il est possible de produire des scores de risque dont la discrimination est comparable à celle des scores de risque de la littérature et dont la calibration est meilleure grâce à l'utilisation de l'algorithme des plus proches voisins. En intégrant dans le processus un spécialiste du domaine et une personne connaissant les besoins des futurs utilisateurs et en y associant l'utilisation d'une méthode de prédiction performante, nous permettons la production de scores de risque pour le cancer du sein qui soient à la fois performants en termes de discrimination et de calibration face aux scores existants, mais également fortement adaptés au contexte afin de maximiser les chances qu'ils soient utilisés.

Pour produire ces scores, nous avons utilisé une base de données publique qui nous permet de confronter nos résultats à ceux d'autres chercheurs et nous avons utilisé une base de données française afin de concevoir un score de risque qui pourra être utilisé en France dans une clinique du risque par exemple. Sur la base publique américaine, avec une combinaison unique de quatre attributs (âge de la femme, densité mammaire, nombre de biopsies et nombre d'antécédents familiaux de cancer du sein) qui affiche une aire sous la courbe ROC de 0,637, nous proposons un score aussi efficace en discrimination, plus efficace en calibration, mais plus simple que les scores disponibles dans la littérature. Sur la base de femmes françaises, nous proposons un score simple et efficace, puisqu'il comprend quatre attributs (âge de la femme, déclaration de maladies bénignes, âge à la ménopause et nombre d'antécédents familiaux de cancer du sein) affichant une calibration quasi parfaite et une aire sous la courbe ROC de 0,625 sans utilisation de la densité mammaire connue pour permettre une amélioration des performances.

Enfin, nous avons montré un prototype d'application web permettant l'utilisation des scores de risque d'une manière plus efficace que les outils qui existent actuellement en mettant l'accent sur l'appropriation de l'évolution du score de risque par l'utilisateur via des composants graphiques qui permettent d'afficher le niveau de risque de l'utilisateur de manière instantanée.

Ces travaux seront très rapidement utilisés dans le cadre de la mise en place d'une clinique du risque à l'Institut Gustave Roussy. Dans un premier temps, le score de risque produit sera intégré à un logiciel permettant de diffuser une information précise du point de vue pratique et scientifique. D'une part, les femmes se rendant à la clinique du risque de l'IGR pourront mesurer leur propre risque de cancer du sein, le comparer avec le risque moyen des participantes à la cohorte E3N et obtenir des informations précises sur les actions de dépistage à mener. D'autre part, l'utilisation d'un tel logiciel permettra aux femmes de s'approprier l'impact des différents facteurs de risque sur la maladie, en utilisant une interface graphique agréable et active. Dans un second temps, le processus proposé permettra de concevoir des modèles d'évaluation du risque qui pourront être utilisés directement par les médecins grâce à l'utilisation de bases de données spécifiques.

Perspectives de recherche

Notre contribution constitue une proposition globale et cohérente permettant la prise en compte du contexte d'utilisation dans la production de scores de risque dans le domaine de la santé grâce à l'utilisation de l'algorithme des plus proches voisins. Il reste cependant des pistes à explorer pour améliorer notre proposition.

Le processus proposé possède deux atouts que sont le mode de choix des attributs intégrés dans la combinaison à l'origine du score de risque et une architecture indépendante de la maladie. En revanche, ce processus pourra être amélioré afin de proposer une étape de choix de la méthode de modélisation qui permette d'adapter ce choix aux besoins et aux possibilités offertes par les données.

Nous avons utilisé un jeu de données construit sur une période de temps fixe qui ne permet pas d'exploiter l'ensemble des données recueillies au fil des années au sein d'une cohorte de grande taille. La prise en compte de la dimension temps dans les données est par conséquent une perspective de travail intéressante afin de tirer le meilleur parti des données de cohorte disponibles dans le monde de la santé.

Au niveau de l'algorithme des plus proches voisins, nous nous sommes limités à l'utilisation d'un faible nombre d'attributs en raison de la nécessité de proposer des scores simples. Mais l'intégration d'un nombre significativement plus élevé d'attributs pourrait poser le problème du temps de calcul nécessaire à la génération de milliers de modèles de risque en fonction des configurations testées et en fonction des attributs dont on veut mesurer la performance. La mise à disposition de données génomiques issues de la cohorte E3N risque en effet de multiplier le nombre d'attributs à disposition.

De plus, des outils ont été conçus sous la forme de feuilles de calcul personnalisées pour permettre la navigation dans les milliers de modèles générés. Cette méthode de parcours des résultats pourrait être améliorée pour permettre de faciliter , d'une part, la gestion des expériences menées et, d'autre part, l'exploration des résultats.

Enfin, nous avons comme perspective de tester rapidement le processus proposé,

associé à l'algorithme des plus proches voisins, sur d'autres maladies étudiées au sein de cohortes différentes. En effet, nous souhaitons confronter l'algorithme à des données dont les valeurs sont distribuées différemment de celles utilisées pour le cancer du sein. La confrontation du processus et de l'architecture du système d'information à d'autres épidémiologistes, d'autres données et d'autres besoins permettra de définir de nouvelles pistes d'amélioration.

Glossaire

A – *Activité* : Partie de traitement réalisée dans un processus métier, qui, d'une part, est définie par les informations qu'elle utilise en entrée et qu'elle produit en sortie, par les événements métier reçus et produits et qui, d'autre part, respecte certaines règles : l'activité métier ne met en œuvre qu'une action principale, elle n'est pas interruptible [Simonin 09].
- *Adipeux* : De nature graisseuse.
- *Allèle* : Un gène est une unité d'information génétique qui, en fonction des mutations, peut exister en plusieurs versions qu'on appelle allèles.
- *Anamnèse* : Récit de l'histoire d'une maladie et de ses circonstances d'un patient à son médecin. Elle constitue un élément de l'examen médical.
- *Anatomo-pathologie* : Science qui a pour objet l'étude des lésions organiques rencontrées au cours des maladies [Brumpt 10]. Plus précisément, c'est l'étude des lésions macroscopiques et microscopiques des tissus pathologiques prélevés sur un sujet. Dans le cas du cancer du sein, le compte-rendu anatomo-pathologique synthétise les différentes caractéristiques du cancer (stade, classification, type histologique, récepteurs hormonaux, etc.).
- *Apoptose* : Processus normal qui constitue la mort programmée de cellules d'un être vivant.
- *Atypie* : Voir hyperplasie atypique.

B – *Biopsie* : Prélèvement d'un échantillon de tissu. Une biopsie mammaire permet de déterminer si une tumeur est bénigne ou maligne.

C – *Calibration* : Capacité d'un score à prédire un nombre de cas de maladie proche du nombre réel de cas de maladie observés (voir page 54).
- *Cas d'utilisation* : Représentation d'une séquence d'actions d'un système faisant intervenir un ou plusieurs acteurs et produisant un résultat mesurable. Correspond à une grande fonctionnalité d'un système. Un système est caractérisé par l'ensemble de ses cas d'utilisation [Simonin 09].
- *Cas-témoin* : À propos d'une étude, voir « étude cas-témoin ».
- *Climatère* : Ensemble des symptômes causés par la ménopause.
- *Cohorte (prospective, rétrospective)* : Ensemble d'individus suivis sur une même période de temps. Une cohorte est dite prospective si les informations sont recueillies sur les individus avant l'occurrence d'une maladie (sur la base de questionnaires par exemple). Elle est dite rétrospective si les informations sont recueillies après l'occurrence de la maladie (sur la base de dossiers médicaux par exemple).
- *Combinaison* : On appelle combinaison dans ce manuscrit, un groupe d'attributs (ou variables ou facteurs de risque) à partir duquel on estimera le risque. Par exemple, une combinaison peut être : âge, âge aux premières règles et âge

au premier accouchement.

- *Composant applicatif* : Réalisation de fonctions présentant une finalité fonctionnelle ou une finalité technique décrite dans des cas d'utilisation. Un composant applicatif produit des données physiques fournies à d'autres composants organiques via ses interfaces. Un composant applicatif peut être l'agrégation de plusieurs composants applicatifs [Simonin 09].
- *Concordance* : Dans le domaine des scores de risque, la concordance correspond à la comparaison des classements qui résultent de l'attribution du niveau de risque prédit par deux scores différents à une même population (voir page 56).
- *Configuration* : On appelle configuration dans ce manuscrit, l'ensemble des paramètres qui permettent de spécifier le fonctionnement d'un algorithme. Par exemple, une configuration pour l'algorithme des plus proches voisins peut être : 4 000 voisins, similarité mesurée par distance euclidienne et validation croisée à quatre couches.
- *Couverture informatique* : On appelle couverture informatique, le périmètre applicatif (appelé aussi système informatique) qui réalise tout ou partie de la vue fonctionnelle du système d'information.

D

- *Discrimination* : Capacité d'un score à séparer les individus malades des individus sains en attribuant des scores élevés à la première catégorie et des scores moins élevés à la seconde (voir page 50).
- *Donnée fonctionnelle* : Objet utilisé ou produit par les îlots* fonctionnels. C'est un objet relatif aux entités* définies par l'utilisation d'un système [Simonin 09].

E

- *Endogène* : Qui est dû à une cause interne. Par exemple, une hormone est dite endogène si elle a été sécrétée par le corps lui-même.
- *Entité* : Objet utilisé ou produit par les cas d'utilisation. C'est soit une notion métier, décrite au niveau de détail nécessaire et suffisant pour les besoins du système, soit une notion propre au système modélisé, dont l'existence est généralement due à la prise en compte de l'organisation [Simonin 09].
- *Épidémiologie* : Discipline qui a pour objet l'étude de l'influence de divers facteurs sur les maladies, notamment sur leur fréquence, distribution et étiologie*.
- *Épithélium* : Tissu composé de cellules jointives recouvrant la surface du corps ou tapissant l'intérieur de certaines cavités naturelles de l'organisme [AF 35].
- *Étiologie* : Discipline qui a pour objet la recherche des causes. Dans le domaine médical, il s'agit de la recherche des causes d'une maladie.
- *Étude cas-témoin* : Une étude cas-témoin consiste à comparer la fréquence d'exposition antérieure à un facteur de risque dans un groupe de malades et dans un groupe de témoins n'ayant pas cette maladie. Les témoins sont choisis pour être représentatifs de la population dont sont issus les cas [Bouyer 95].
- *Exogène* : Qui est dû à une cause externe. Par exemple, une hormone est dite exogène si elle provient d'un traitement médical.

F

- *Facteur de confusion* : Facteur qui présente une association avec l'exposition ou le facteur de risque examiné et qui peut influencer le résultat. Un facteur de confusion peut affaiblir ou renforcer une association entre l'exposition et

les résultats observés. De ce fait, un lien inexistant dans la réalité peut être suggéré, ou, au contraire, un lien réel peut être méconnu [Minerva 08].
- *Fil d'exécution* : Séquence d'instructions informatiques exécutée par le processus d'un ordinateur.

G – *Germinale* : Se dit des cellules qui peuvent être à l'origine des gamètes (dont les ovules et spermatozoïdes). S'oppose aux cellules somatiques*.

H – *Hormone* : Substance sécrétée par le système endocrinien qui permet de transporter un message par le sang vers une cellule qui possède des récepteurs spécifiques.
- *Hyperplasie atypique* : L'hyperplasie est une prolifération de cellules. Elle peut être simple ou atypique si la prolifération est anormale.

I – *Îlot fonctionnel* : Regroupement de parcelles* fonctionnelles qui est le plus bas niveau de découpe du Plan Local d'Urbanisme possédant une vue externe [Simonin 09].
- *Incidence* : Nombre de nouveaux cas d'une maladie dans une population donnée de non-malades et sur une période de temps donnée. En épidémiologie, l'unité la plus fréquemment utilisée pour mesurer le taux d'incidence est le nombre de « personnes-années ». À différencier de la prévalence*.
- *« include »* : Un cas d'utilisation inclut un autre cas d'utilisation si le déroulement du premier induit le déroulement du second.
- *Indice de Masse Corporelle (IMC)* : Rapport de la masse corporelle (en kilogrammes) sur la taille (en mètres) au carré. Défini par l'Organisation Mondiale de la Santé, cet indice permet d'évaluer la corpulence et les risques liés au surpoids.
- *Interface applicative* : Vue externe d'un composant applicatif offrant un accès à des données physiques ou à des traitements sur ces données [Simonin 09].
- *Investigateur* : Nom donné au responsable scientifique d'une étude de cohorte*.

J – *Java* : Langage de programmation orientée objet maintenu par la société Oracle.

L – *Lien de communication* : Permet le transfert d'information entre nœuds d'exécution [Simonin 09].

M – *Machine virtuelle* : Système d'exploitation invité qui est exécuté par le système d'exploitation hôte en charge des relations avec la partie physique de l'ordinateur. Les machines virtuelles permettent d'utiliser simultanément plus d'un système d'exploitation par machine physique.
- *Mammographie* : Examen radiographique du sein.
- *Ménarche* : Apparition des premières menstruations chez la femme.
- *Méta-analyse* : En épidémiologie, une méta-analyse consiste en la comparaison statistique de multiples études sur un même thème afin de mettre en évidence des points communs sur les origines d'une maladie ou d'expliquer les points divergents entre les études.
- *Mitose* : Synonyme de division cellulaire. Désigne la phase de séparation des chromosomes sur le point de créer deux cellules filles à partir d'une cellule

mère.

N – *Nullipare* : Se dit d'une femme qui n'a jamais accouché.

– *Nœud d'exécution* : Machine physique ou environnement d'exécution permettant de supporter l'exécution des systèmes [Simonin 09].

P – *Paramétrique (modèle)* : Une méthode de modélisation est dite paramétrique si elle repose sur l'hypothèse que les données suivent une distribution particulière ou, de manière plus générale, si les données ne suivent pas une structure de modèle fixé. Par opposition, une méthode de modélisation non-paramétrique ne repose pas sur une hypothèse de distribution définie a priori.

– *Parcelle fonctionnelle* : Fonction élémentaire qui est le niveau de description fonctionnelle le plus bas d'un îlot* fonctionnel [Simonin 09].

– *Parité* : Nombre d'enfants mis au monde vivants par une femme.

– *Polygénique* : En référence à l'origine d'une maladie : signifie que plusieurs gènes en sont la cause.

– *Prévalence* : Nombre de cas d'une maladie dans une population donnée à un instant donné. À différencier de l'incidence*.

– *Prévention primaire* : Mesures utilisées pour tenter d'éviter la survenue de la maladie dans une population de personnes non-malades et à faible risque (exemple : vaccination ou conseils nutritionnels type Plan National Nutrition Santé : « Cinq fruits et légumes par jour »).

– *Processus métier* : Séquence d'actes réalisée par l'entreprise qui produit un résultat dont la valeur est perceptible et mesurable pour un acteur individuel de l'entreprise. Il doit être défini indépendamment de toute organisation et de tout système existant dans l'entreprise [Simonin 09].

– *Profil* : On appelle profil dans ce manuscrit, les valeurs prises par les attributs (ou variables, ou facteurs de risque) qui décrivent la femme. Un profil de femme peut être par exemple : 46 ans, premières règles à 12 ans, premier accouchement à 28 ans.

R – *Risque relatif* : Mesure permettant d'évaluer le risque de survenue d'un événement entre deux populations. Par exemple, si 10 % des femmes ayant un parent atteint par un : cancer du sein sont elles mêmes touchées par un cancer du sein alors que 5 % des femmes qui n'ont pas de parent atteint sont touchées par un cancer du sein, alors le risque relatif vaut 10/5, soit 2. Les femmes qui ont un parent atteint par un cancer du sein ont donc deux fois plus de risque d'être atteintes par un cancer du sein.

– *Rôle métier* : Représentation jouée dans la réalisation d'un processus* métier par un individu ou une unité organisationnelle, interne au domaine modélisé (par exemple toute l'entreprise). Un rôle métier contribue à la réalisation d'un ou plusieurs processus métier et plusieurs rôles métier peuvent contribuer à la réalisation d'un processus métier. Un rôle métier émet et reçoit des événements métier. [Simonin 09].

S – *Sensibilité* : Dans le domaine médical, capacité d'un test à donner un résultat positif lorsque la maladie est présente.

- *Somatique* : Par opposition aux cellules germinales*, les cellules somatiques sont celles qui ne seront pas à l'origine des gamètes.
- *Spécificité* : Dans le domaine médical, capacité d'un test à donner un résultat négatif lorsque la maladie est absente.

T
- *Tâche* : Partie de traitement réalisée dans une procédure métier qui, d'une part est définie par les informations (données métier) qu'elle utilise en entrée et qu'elle produit en sortie, par les événements métier reçus et produits et qui, d'autre part, respecte certaines règles : l'activité métier ne met en œuvre qu'une action principale, elle n'est pas interruptible et elle est réalisée par un rôle métier unique à un moment donné [Simonin 09].

V
- *Validité externe* : À propos d'un score : consiste à déterminer ses qualités lorsqu'il est appliqué à un autre échantillon de la même population (synonyme : reproductibilité).
- *Validité interne* : À propos d'un score : consiste à déterminer ses qualités lorsqu'il est appliqué à l'échantillon qui a permis de le mettre au point.

Abbréviations

- ARC : Association pour la Recherche sur le Cancer
- ATM : Ataxia Telangiectasia Mutated
- AUC : Area Under the Curve (Aire sous la courbe)
- BCSC : Breast Cancer Surveillance Consortium (Consortium pour de dépistage du cancer du sein)
- BRCA : BReast CAncer (Nom d'une mutation génétique)
- CEPI-DC : Centre d'ÉPIdémiologie sur les causes médicales de DéCès
- CNIL : Commission Nationale de l'Informatique et des Libertés
- CNRS : Centre National de la Recherche Scientifique
- CRISP-DM : CRoss Industry Standard Process for Data Mining
- E3N : Étude Épidémiologique auprès des femmes de la MGEN
- HDL : High Density Lipoproteins
- HTMl : HyperText Markup Language (Langage à balises pour l'hypertexte)
- IGR : Institut Gustave Roussy
- IMC : Indice de Masse Corporelle
- INSERM : Institut National de la Santé et de la Recherche Médicale
- IRM : Imagerie par Résonnance Magnétique
- NCI : National Cancer Institute
- OMS : Organisation mondiale de la santé
- PHP : PHP : Hypertext Preprocessor
- PLU : Plan Local d'Urbanisme
- ROC : Receiver Operating Characteristic (Fonction d'efficacité du récepteur)
- SEER : Surveillance Epidemiology and End Results

Tableau 5.20 : Tableau des abbréviations d'attributs utilisés dans le manuscrit pour le cancer du sein

Nom abrégé	Attribut
age	Âge de la femme
agemeno	Âge de la femme à la ménopause
statmeno	Statut ménopausique
typemeno	Type de ménopause (naturelle, artificielle)
menarche	Âge de la femme à la ménarche
mbs	Nombre de maladies bénignes du sein
agenai	Âge de la femme à la naissance du premier enfant
biop	Nombre de biopsies pratiquées
kdeg1	Nombre d'antécédents au premier degré
dens	Densité mammaire
precmamm	Résultat de la précédente mammographie
imc	Indice de masse corporelle
ths	Traitement hormonal substitutif de la ménopause

A

Annexes – Documents d'architecture

A.1 Cas d'utilisation

A.1.1 Cas d'utilisation CuSélectionAttributs

Résumé : Ce cas d'utilisation permet de décrire l'étape de sélection des attributs, ou facteurs de risque, dont la capacité prédictive sera testée au sein d'un score de risque.

Contexte de déclenchement : Le cas d'utilisation peut être déclenché par la conception d'un nouveau modèle (voir CuConceptionModèle).

Rôles : Expert en fouille de données

Pré-conditions : Une base de données épidémiologiques est disponible. Un processus de conception de modèle est en cours.

Description : Après déclenchement du cas d'utilisation, l'expert en fouille de données construit la liste des attributs à disposition dans la base de données et qui ont un lien avec la maladie étudiée.

Post-conditions : Une liste d'attributs sélectionnée est constituée

Exception : Les données n'existent pas sur le serveur de stockage.

Scénarios : Scénario nominal de sélection des attributs à utiliser :

ScSélectionAttributs
1. Sélection d'un attribut dans la base de données
2. Vérification du lien entre la maladie et l'attribut sélectionnée
3. Ajout de l'attribut à la liste

A.1.2 Cas d'utilisation CuFiltrageAttributs

Résumé : Ce cas d'utilisation permet de décrire l'étape de filtrage de la liste d'attributs par l'épidémiologiste.

Contexte de déclenchement : Le cas d'utilisation est déclenché lors de la conception d'une nouvelle liste d'attributs par spécialité après la première phase de sélection des attributs par l'expert en fouille de données.

Rôles : Épidémiologiste

Pré-conditions : Une liste d'attributs sélectionnés est disponible.

Description : Après déclenchement du cas d'utilisation, l'épidémiologiste filtre la liste des attributs en fonction de ses connaissances. Le détail de ce cas d'utilisation est donné page 76.

Post-conditions : Une liste d'attributs filtrée est disponible.

Exception : La liste des attributs de départ n'existe pas ou est vide. Tous les attributs sont éliminés.

Scénarios : Scénario nominal de filtrage des attributs à utiliser :

ScFiltrageAttributs
1. Boucle sur les attributs disponibles, pour chaque attribut
2. Étude d'un attribut de la liste de départ
3. Choix de sa conservation ou de son élimination
4. Mise à jour de son statut validé/pas-validé

A.1.3 Cas d'utilisation CuValidationAttributs

Résumé : Ce cas d'utilisation permet de décrire l'étape de validation par l'analyste de la liste d'attributs filtrée.

Contexte de déclenchement : Le cas d'utilisation est déclenché à la suite du filtrage d'une nouvelle liste des attributs.

Rôles : Analyste du besoin

Pré-conditions : Une liste d'attributs filtrée est disponible.

Description : Après déclenchement du cas d'utilisation, l'analyste valide la liste des attributs selon ses connaissances du contexte d'utilisation. Le détail de ce cas d'utilisation est donné page 76.

Post-conditions : Une liste d'attributs validés est disponible.

Exception : La liste d'attributs filtrée n'existe pas ou est vide. L'analyste ne valide pas la liste des attributs.

Scénarios : Scénario nominal de validation des attributs à utiliser :

ScValidationAttributs
1. Boucle sur les attributs filtrés de la liste des attributs
2. Choix de la validation ou non de l'attribut
3. Si tous les attributs sont validés, la liste est validée

A.1.4 Cas d'utilisation CuPréparationDonnées

Résumé : Ce cas d'utilisation permet de décrire l'étape de préparation des données pour, d'une part les mettre dans un format compatible avec l'algorithme retenu et, d'autre part, permettre la meilleure discrétisation possible.

Contexte de déclenchement : Le cas d'utilisation peut être déclenché par la conception d'un nouveau modèle.

Rôles : Expert en fouille de données

Pré-conditions : Une base de données épidémiologique est disponible. L'algorithme de modélisation qui sera utilisé est défini. La liste validée des attributs est disponible. Les règles de gestion des valeurs manquantes et de discrétisation sont disponibles.

Description : En suivant des règles de décisions, les données sont préparées conformément au processus décrit page 77.

Post-conditions : Une liste de données préparées est prête à être utilisée.

Exception : Il n'existe pas de liste validée des attributs. Il n'existe pas de règles de décision disponibles. Les données ne peuvent pas être nettoyées (trop de données manquantes par exemple)

Scénarios : Scénario nominal de préparation des données :

ScPréparationDonnées
1. Extraction des données depuis le serveur de stockage
2. Gestion des données manquantes
3. Discrétisation des données
4. Exportation des données dans un format utilisable par l'algorithme choisi

A.1.5 Cas d'utilisation CuValidationPréparation

Résumé : Ce cas d'utilisation permet de décrire l'étape de validation par l'épidémiologiste de la préparation effectuée sur les données.

Contexte de déclenchement : Le cas d'utilisation est déclenché à la suite de la préparation des données.

Rôles : Épidémiologiste

Pré-conditions : Une préparation des données est disponible.

Description : Après déclenchement du cas d'utilisation, l'épidémiologiste valide la préparation des données selon ses connaissances du domaine et les règles de décisions qui ont été utilisées pour préparer les données. Le détail de la description de cas d'utilisation correspond au processus décrit page 77.

Post-conditions : La préparation des données est validée.

Exception : Les données préparées ne sont pas disponibles. L'épidémiologiste ne valide pas la préparation des données.

Scénarios : Scénario nominal de validation de la préparation des données à utiliser :

ScValidationPréparation
1. Étude de chacun des attributs préparés
2. Si tous les attributs sont validés, la préparation est validée

A.1.6 Cas d'utilisation CuCalculPerformances

Résumé : Calcul de la performance de chacune des combinaisons d'algorithmes, attributs, répartitions possibles par la génération d'indicateurs de discrimination, d'indicateurs de calibration et de statistiques descriptives sur l'utilisation des algorithmes.

Contexte de déclenchement : Le cas d'utilisation est déclenché par la génération d'un ou plusieurs modèles.

Rôles : Expert en fouille de données

Pré-conditions : Une liste de modèles de risque est disponible.

Description : La description de ce cas d'utilisation correspond au processus décrit page 77.

Post-conditions : Une liste des différentes mesures de performance est disponible. Les tables de référence des risques sont conservées.

Exception : La puissance de calcul n'est pas disponible ou en trop faible quantité. L'espace mémoire est trop faible au vu du nombre de résultats.

Scénarios : Scénario nominal de calcul des performances :

ScCalculPerformances
1. Définition des différentes configurations d'algorithme à tester
2. Définition des différentes combinaisons d'attributs de score à tester
3. Définition des différents jeux d'apprentissage/validation à utiliser
4. Génération d'un modèle pour chaque paire de jeu/combinaison
5. Calcul d'indicateurs de performance pour chaque algorithme, configuration, combinaison, répartition défini
6. Exportation des résultats (modélisations et leurs indicateurs associés)

A.1.7 Cas d'utilisation CuChoixModèle

Résumé : Choix d'un modèle (association d'un algorithme, d'une configuration d'algorithme, d'une composition en attributs et d'une répartition des jeux d'apprentissage/validation) parmi ceux testés en fonction de différents critères, dont les performances, la connaissance de la spécialité médicale et le contexte.

Contexte de déclenchement : Le cas d'utilisation est déclenché par la conception de nouveaux modèles.

Rôles : Analyste du besoin ou épidémiologiste

Pré-conditions : Une liste de modèles est disponible avec ses descripteurs de la performance associée.

Description : Après déclenchement du cas d'utilisation, l'analyste ou l'épidémiologiste sélectionne un modèle en fonction du contexte d'utilisation futur du score. La description complète de ce cas d'utilisation est donnée en page 77.

Post-conditions : Un modèle de risque est choisi.

Exception : Une liste de mesures de performance n'est pas disponible. Aucun choix n'est effectué par le sélectionneur.

Scénarios : Scénario nominal de choix du modèle :

ScChoixModèle
1. Obtention des performances des différents modèles testés
2. Évaluation des performances obtenues
3. Choix d'un modèle

A.1.8 Cas d'utilisation CuValidationChoixModèle

Résumé : Ce cas d'utilisation permet de décrire l'étape de validation par l'épidémiologiste ou l'analyste du choix du modèle effectué.

Contexte de déclenchement : Le cas d'utilisation est déclenché à la suite de la génération des différents modèles.

Rôles : Épidémiologiste (si l'analyste a choisi le modèle) ou l'analyste (si l'épidémiologiste a choisi le modèle)

Pré-conditions : Un choix de modèle a été effectué.

Description : Après déclenchement du cas d'utilisation, l'épidémiologiste ou l'analyste valide le choix du modèle effectué auparavant. La description complète de ce cas d'utilisation est donnée en page 77.

Post-conditions : Le choix du modèle est validé.

Exception : Une liste de modèles n'est pas disponible. Aucun modèle n'est validé.

Scénarios : Scénario nominal de validation des attributs à utiliser :

ScValidationChoixModèle
1. Étude de chacun des attributs du modèle
2. Étude des indicateurs de performance du modèle
3. Si tous les attributs et les indicateurs sont validés, le modèle est validé

A.1.9 Cas d'utilisation CuGénérerTableRéférence

Résumé : Ce cas d'utilisation permet de décrire l'étape de génération de la table de référence.

Contexte de déclenchement : Le cas d'utilisation est déclenché à la suite de la validation d'un modèle choisi.

Rôles : Expert en fouille de données

Pré-conditions : Un modèle a été choisi et validé.

Description : Une table de référence comprenant les scores précalculés est générée. Le détail de la description de ce processus est disponible page 77.

Post-conditions : Une table de référence est générée.
Exception : Un modèle validé n'est pas disponible.
Scénarios : Scénario nominal de génération d'une table de référence :

ScGénérationTableRéférence
1. Parcours de tous les profils d'individus potentiellement rencontrables
2. Calcul du score de risque pour l'individu
3. Mise en mémoire dans une table de référence
4. Exportation de la table de référence

A.1.10 Cas d'utilisation CuSaisirProfil

Résumé : Ce cas d'utilisation permet de décrire la phase de saisie du profil de l'utilisateur du score de risque.

Contexte de déclenchement : Le cas d'utilisation est déclenché lorsqu'une personne souhaite obtenir son niveau de risque pour la maladie dont le risque a été modélisé.

Rôles : Utilisateur du système d'évaluation du niveau de risque

Pré-conditions : Le risque de maladie dont l'utilisateur veut obtenir le niveau doit avoir été modélisé.

Description : L'utilisateur du système d'évaluation du niveau de risque saisit son profil sur l'interface mise à sa disposition. Les détails de ce cas d'utilisation sont donnés page 76.

Post-conditions : Le profil de l'utilisateur du système d'évaluation du niveau de risque est saisi.

Exception : Aucune.

Scénarios : Non applicable.

A.1.11 Cas d'utilisation CuObtenirRisque

Résumé : Ce cas d'utilisation permet de décrire l'obtention d'un niveau de risque pour un profil d'utilisateur donné et une maladie donnée.

Contexte de déclenchement : Le cas d'utilisation est déclenché lorsqu'une personne veut évaluer son risque de maladie après saisie de son profil.

Rôles : Utilisateur du système d'évaluation du niveau de risque

Pré-conditions : Le profil à évaluer est saisi.

Description : L'utilisateur demande au système l'évaluation d'un risque.
Le système propose à l'utilisateur une liste des attributs.
L'utilisateur renseigne les valeurs de chaque attribut.
Le système renvoie une évaluation du risque par rapport à une table de référence (voir CuConceptionModèle)

Post-conditions : Un niveau de risque est fourni.

Exception : Le profil à évaluer n'est pas saisi. La table de référence ne contient pas le niveau de risque du profil à évaluer.

Scénarios : Scénario nominal d'obtention d'un niveau de risque :

ScObtenirNiveauRisque
1. Demande d'obtention d'un niveau de risque
2. Identification des caractéristiques du profil à évaluer
3. Vérification de la disponibilité du niveau de risque
4. Obtention du niveau de risque du profil
5. Envoi du niveau de risque du profil

A.2 DIAGRAMMES DE SÉQUENCE

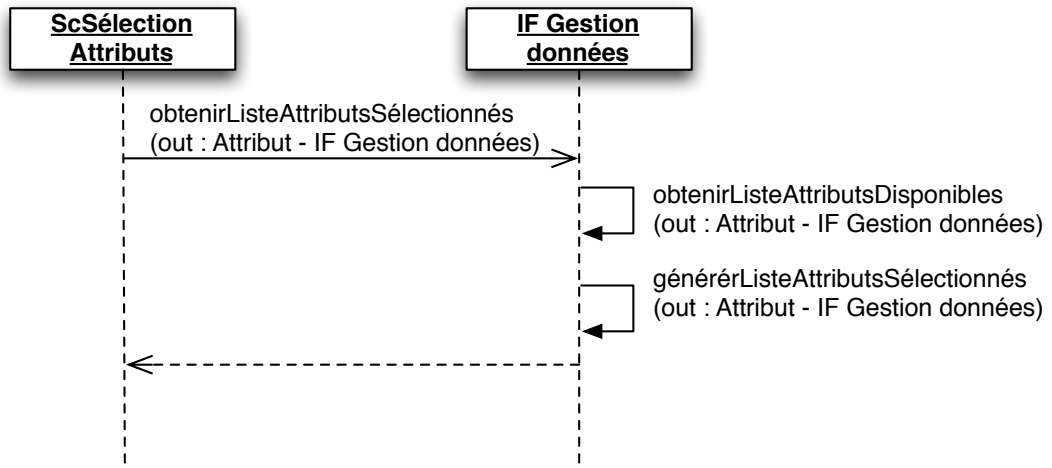

Figure A.1 : Diagramme de séquence fonctionnelle correspondant au scénario ScSélectionAttributs

A

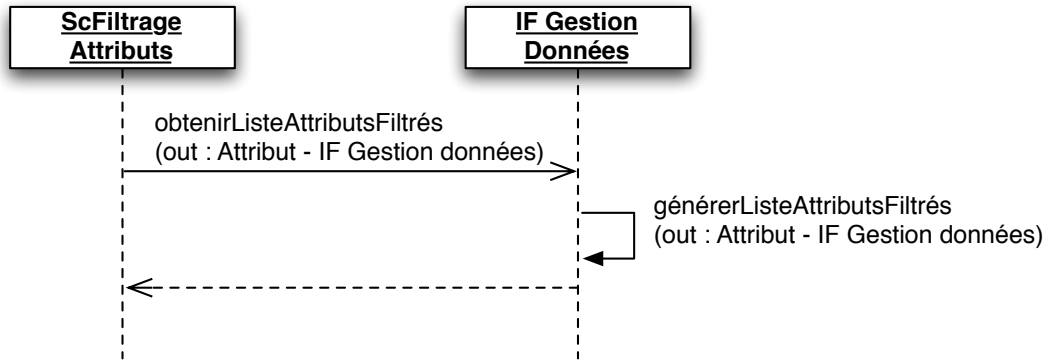

Figure A.2 : Diagramme de séquence fonctionnelle correspondant au scénario ScFiltrageAttributs

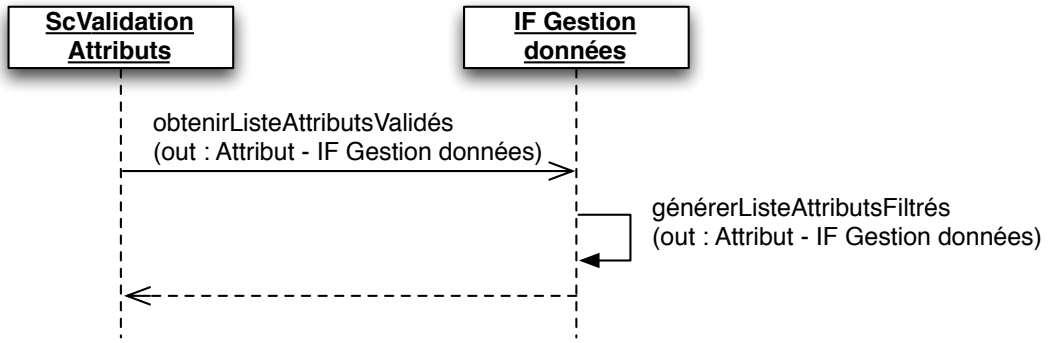

Figure A.3 : Diagramme de séquence fonctionnelle correspondant au scénario ScValidationAttributs

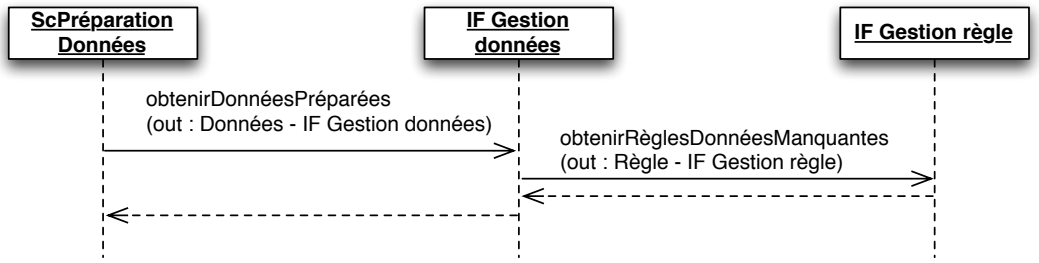

Figure A.4 : Diagramme de séquence fonctionnelle correspondant au scénario ScPréparationDonnées

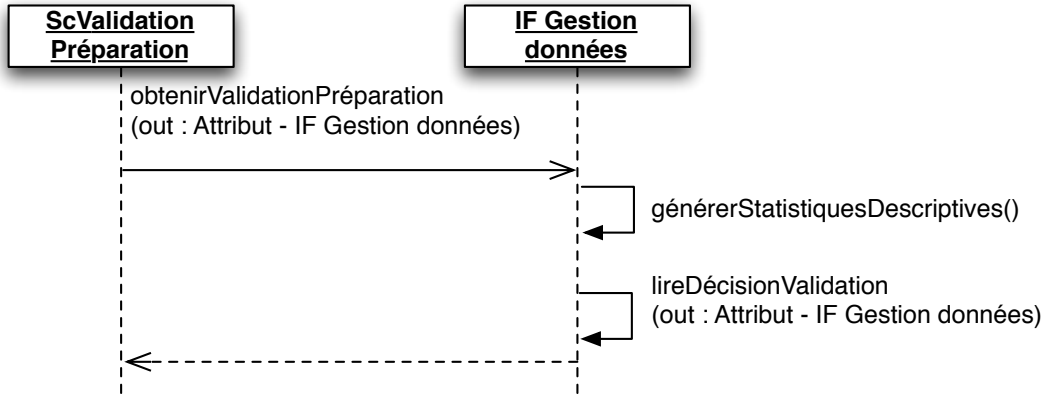

Figure A.5 : Diagramme de séquence fonctionnelle correspondant au scénario ScValidationPréparation

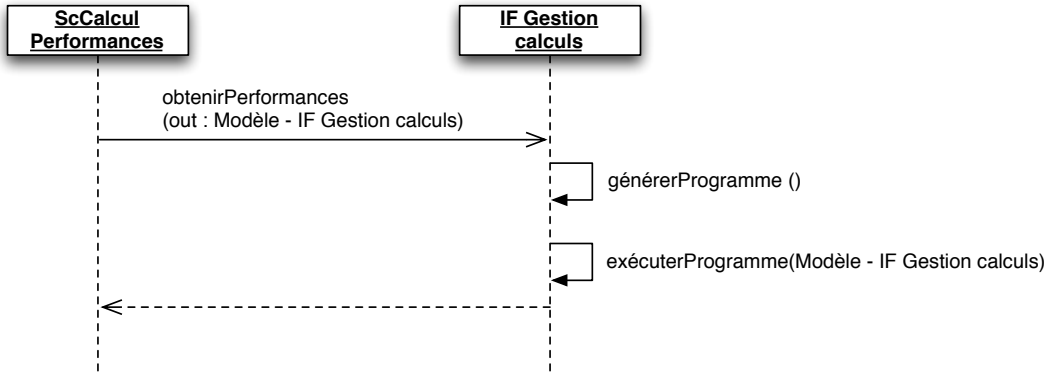

Figure A.6 : Diagramme de séquence fonctionnelle correspondant au scénario ScCalculPerformances

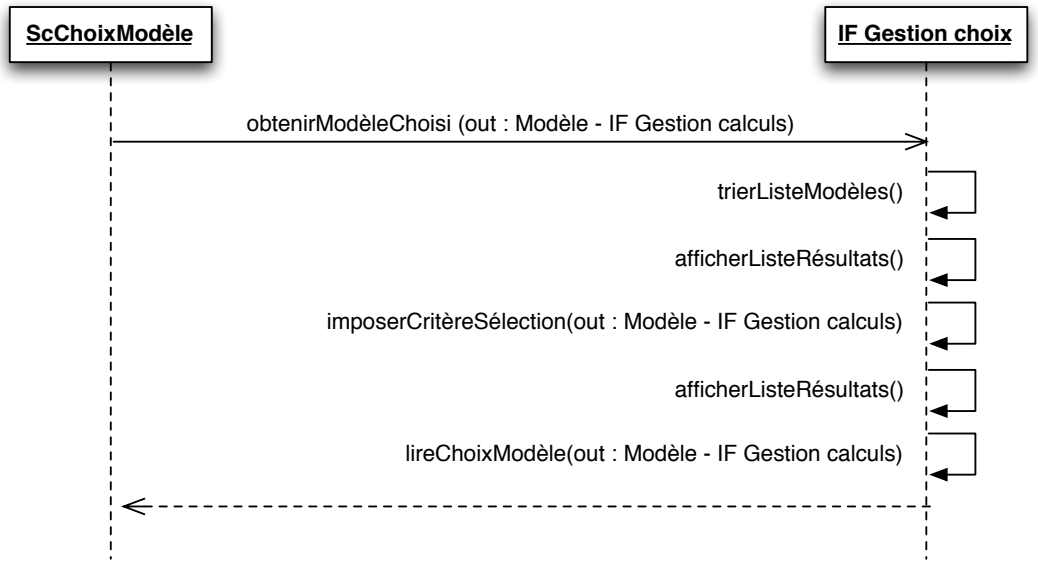

Figure A.7 : Diagramme de séquence fonctionnelle correspondant au scénario ScChoixModèle

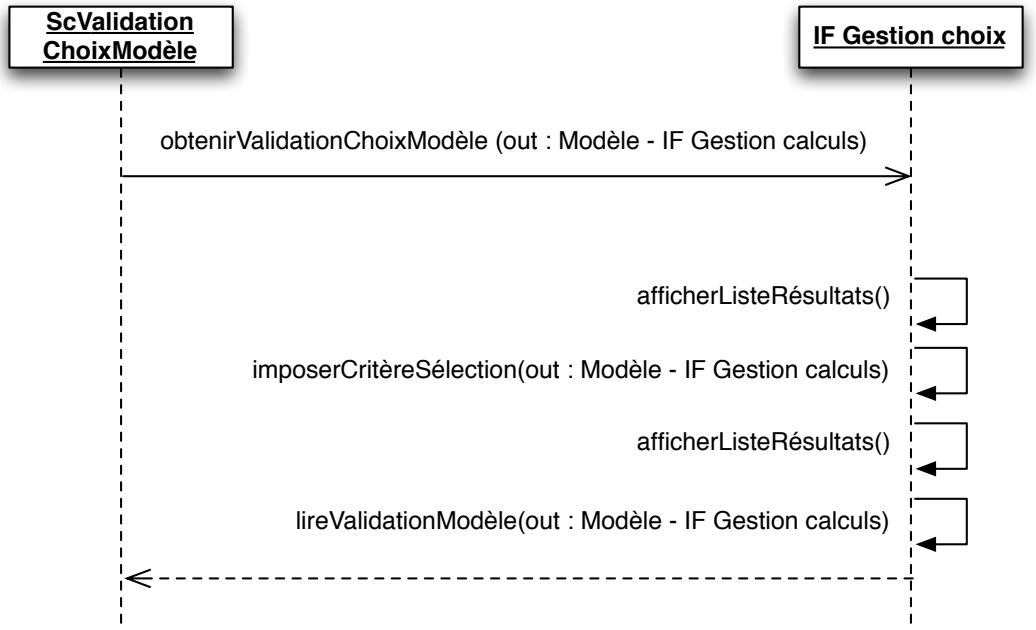

Figure A.8 : Diagramme de séquence fonctionnelle correspondant au scénario ScValidationChoixModèle

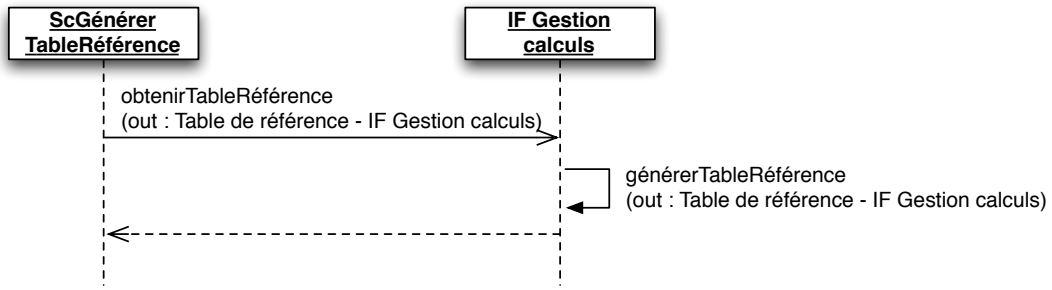

Figure A.9 : Diagramme de séquence fonctionnelle correspondant au scénario ScGénérerTableRéférence

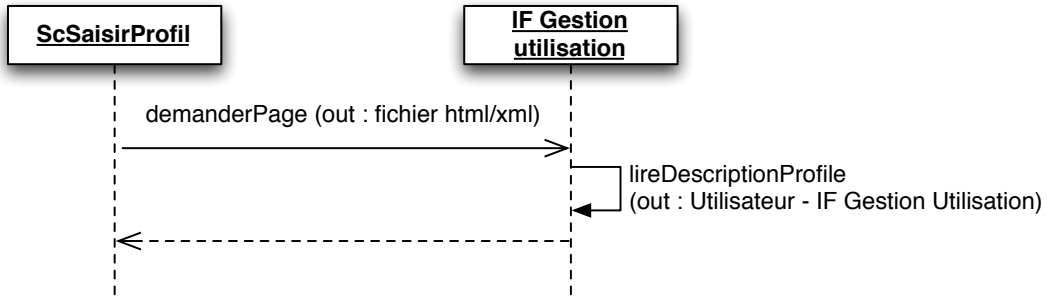

Figure A.10 : Diagramme de séquence fonctionnelle correspondant au scénario ScSaisirProfil

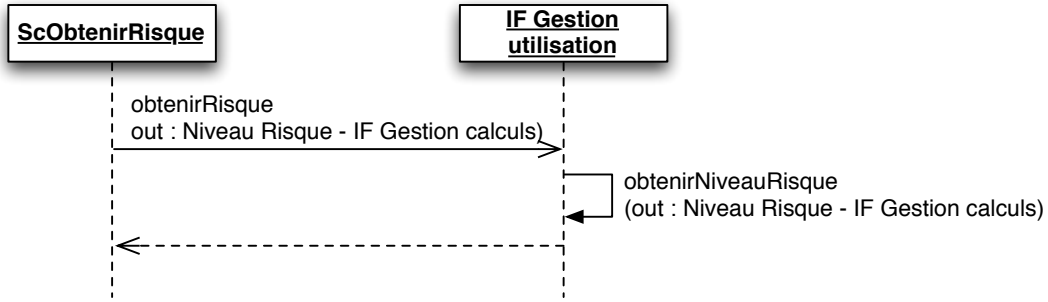

Figure A.11 : Diagramme de séquence fonctionnelle correspondant au scénario ScObtenirRisque

Bibliographie

[AF 35] Académie française AF. Dictionnaire de l'académie française. Librairie Hachette, huitième édition, 1932-1935. 154

[Akoka 07] J. Akoka, L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué, Z. Kedad, S. Nugier, V. Peralta & S. Si-Said Cherfi. *A Framework for Quality Evaluation in Data Integration Systems*. In ICEIS'07 10th Int. Conf. on Enterprise Information Systems, pages 170–175, 2007. 64

[Allred 01] D. C. Allred, S. K. Mohsin & S. A. Fuqua. *Histological and biological evolution of human premalignant breast disease.* Endocrine-Related Cancer, vol. 8, no. 1, pages 47–61, 2001. 16

[ANAÉS 04] Agence Nationale d'Accréditation et d'Évaluation en santé ANAÉS. *Méthodes d'évalusation du risque cardio-vasculaire global*. Rapport technique, Agence Nationale d'Accréditation et d'Évaluation en santé, 2004. 7, 29, 62

[Anderson 91] K.M. Anderson, P.W. Wilson, P.M. Odell & W.B. Kannel. *An updated coronary risk profile. A statement for health professionals*. Circulation, vol. 83, no. 1, pages 356–362, 1991. 25, 28, 29

[Asuncion 07] A. Asuncion & D. J. Newman. *UCI Machine Learning Repository*. http ://www.ics.uci.edu/∼mlearn/MLRepository.html, 2007. 68

[Bach 03] P. B. Bach, M. W. Kattan, M. D. Thornquist, M. G. Kris, R. C. Tate, M. J. Barnett, L. J. Hsieh & C. B. Begg. *Variations in Lung Cancer Risk Among Smokers*. Journal of the National Cancer Institute, vol. 95, no. 6, pages 470–478, 2003. 30

[Ballard-Barbash 97] R. Ballard-Barbash, S.H. Taplin, B.C. Yankaskas, V.L. Ernster, R.D. Rosenberg, P.A. Carney, W.E. Barlow, B.M. Geller, K. Kerlikowske, B.K. Edwards, C.F. Lynch, N. Urban, C.A. Chrvala, C.R. Key, S.P. Poplack, J.K. Worden & L.G. Kessler. *Breast Cancer Surveillance Consortium : a national mammography screening and outcomes database*. American Journal of Roentgenology, vol. 169, no. 4, pages 1001–1008, 1997. 64

[Barlow 06] W. E. Barlow, E. White, R. Ballard-Barbash, P. M. Vacek, L. Titus-Ernstoff, P. A. Carney, J. A. Tice, D. S. M. Buist, B. M. Geller, R. Rosenberg, B. C. Yankaskas & K. Kerlikowske.

Prospective Breast Cancer Risk Prediction Model for Women Undergoing Screening Mammography. Journal of the National Cancer Institute, vol. 98, no. 17, pages 1204–1214, 2006. 36, 65, 82, 121, 124, 132, 134, 137, 145

[Bellaachia 06] A. Bellaachia & E. Guven. *Predicting breast cancer survivability using data mining techniques*. In Proceedings of Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006), 2006. 38, 40

[Belot 08] A. Belot, P. Grosclaude, N. Bossard, E. Jougla, E. Benhamou, P. Delafosse, A.V. Guizard, F. Molinié, A. Danzon, S. Bara, A.M. Bouvier, B. Trétarre, F. Binder-Foucard, M. Colonna, L. Daubisse, G. Hédelin, G. Launoy, N. Le Stang, M. Maynadié, A. Monnereau, X. Troussard, J. Faivre, A. Collignon, I. Janoray, P. Arveux, A. Buemi, N. Raverdy, C. Schvartz, M. Bovet, L. Chérié-Challine, J. Estève, L. Remontet & M. Velten. *Cancer incidence and mortality in France over the period 1980-2005.* Revue d'Épidémiologie et de Santé Publique, vol. 56, no. 3, pages 159–75, 2008. 13

[Bendraou 07] R. Bendraou & M.P. Gervais. *A Framework for Classifying and Comparing Process Technology Domains*. In Proceedings of International Conference on Software Engineering Advances, pages 5–12. IEEE Computer Society Press, 2007. 89

[Beral 03] V. Beral & Collaborative Group. *Breast cancer and hormone-replacement therapy in the Million Women Study*. The Lancet, vol. 362, no. 9382, pages 419–427, 2003. 19

[Booch 04] G. Booch, J. Rumbaugh & I. Jacobson. Unified modeling language reference manual, the (2nd edition). Pearson Higher Education, 2004. 88, 96, 100, 101

[Bouyer 95] J. Bouyer, D. Hémon & S. Cordier. Épidémiologie : Principes et méthodes quantitatives. Inserm, 1995. 154

[Boyd 05] N. F. Boyd, J.M. Rommens, K. Vogt, V. Lee, J. L. Hopper, M.J. Yaffe & A.D. Paterson. *Mammographic breast density as an intermediate phenotype for breast cancer*. The Lancet Oncology, vol. 6, no. 10, pages 798–808, 2005. 19

[Brisson 08] L. Brisson & M. Collard. *An ontology driven data mining process*. In José Cordeiro & Joaquim Filipe, editeurs, ICEIS 2008 : proceedings of the Tenth International Conference on Enterprise Information Systems, June 12-16, Barcelona, Spain, pages 54–61, 2008. http ://www.iceis.org/iceis2008/. 77

[Brisson 09] L. Brisson & M. Collard. How to semantically enhance a data mining process ?, volume 19 of *Lecture Notes in Business Information Processing*, chapitre Enterprise information systems, pages 103 – 116. Springer Berlin Heidelberg, 2009. 77

[Brumpt 10] É. Brumpt. Précis de parasitologie. Collection de précis médicaux. Masson, 1910. 153

[Buell 73] P. Buell. *Changing Incidence of Breast Cancer in Japanese-American Women.* Journal of the National Cancer Institute, vol. 51, no. 5, pages 1479–1483, 1973. 20

[Burstein 04] H. J. Burstein, K. Polyak, J. S. Wong, S. C. Lester & C. M. Kaelin. *Ductal carcinoma in situ of the breast.* The New England Journal of Medicine, vol. 350, no. 14, pages 1430–1441, 2004. 16

[Chajès 08] V. Chajès, C.M. A. Thiébaut, M. Rotival, E. Gauthier, V. Maillard, M.-C. Boutron-Ruault, V. Joulin, M. G. Lenoir & F. Clavel-Chapelon. *Association between serum trans-monounsaturated fatty acids and breast cancer risk in the E3N-EPIC Study.* American Journal of Epidemiology, vol. 167, no. 11, pages 1312–20, 2008. 20

[Chapman 00] P. Chapman, J. Clinton, R. Kerber & T. Khabaza. *CRISP-DM 1.0 Step-by-step data mining guide.* Rapport technique, The CRISP-DM Consortium, 2000. xiii, 72

[Chen 06] J. Chen, D. Pee, R. Ayyagari, B. Graubard, C. Schairer, C. Byrne, J. Benichou & M. H. Gail. *Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density.* Journal of the National Cancer Institute, vol. 98, no. 17, pages 1215–1226, 2006. 145

[Claus 91] E. B. Claus, N. Risch & W. D. Thompson. *Genetic analysis of breast cancer in the cancer and steroid hormone study.* The American Journal of Human Genetics, vol. 48, no. 2, pages 232–242, 1991. xi, 32, 33

[Clavel-Chapelon 96] F. Clavel-Chapelon, C. Jadand, H. Goulard & C. Guibout-Peigné. *E3N, a cohort study on cancer risk factors in MGEN women. Description of protocol, main characteristics and population.* Bulletin du Cancer, vol. 83, no. 12, pages 1008–1013, 1996. 66

[Clavel-Chapelon 97] F. Clavel-Chapelon, M.J. van Liere, C. Giubout, M.Y. Niravong, H. Goulard, C. Le Corre, L.A. Hoang, J. Amoyel, A. Auquier & E. Duquesnel. *E3N, a French cohort study on cancer risk factors. E3N Group. Etude Epidémiologique auprès de*

femmes de l'Education Nationale. European Journal of Cancer Prevention, vol. 6, no. 5, pages 473–478, 1997. 66

[Clavel-Chapelon 02] F. Clavel-Chapelon & M. Gerber. *Reproductive factors and breast cancer risk. Do they differ according to age at diagnosis ?* Breast Cancer Research and Treatment, vol. 72, no. 2, pages 107–115, 2002. 19

[CNIL 12] CNIL. *Commission Nationale de l'Informatique et des Libertés.* http ://www.cnil.fr/en-savoir-plus/fiches-pratiques/fiche/article/un-imperatif-la-securite/, 2012. 70

[Cohen 83] S. Cohen, T. Kamarck & R. Mermelstein. *A global measure of perceived stress.* Journal of Health and Social Behavior, vol. 24, pages 385–396, 1983. 24

[Collishaw 09] N.E. Collishaw, N.F. Boyd, K.P. Cantor, S.K. Hammond, K.C. Johnson, J. Millar, A.B. Miller, M. Miller, J.R. Palmer, A.G. Salmon & F. Turcotte. *Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk.* Rapport technique, Ontario Tobacco Research Unit, 2009. 20

[Comyn-Wattiau 10] I. Comyn-Wattiau, J. Akoka & L. Berti-Equille. *La qualité des systèmes d'information – Vers une vision plus intégrée.* Ingénierie des Systèmes d'Information, vol. 15, no. 6, pages 9–32, 2010. 87

[Cong 11] C. Cong & C. P. Tsokos. *Parametric and Nonparametric Analysis of Breast Cancer Treatments.* International Journal of Biological & Life Sciences, vol. 7, no. 3, pages 134–137, 2011. 38, 40

[Consortium 06] Breast Cancer Surveillance Consortium. Site internet, 2006. 121

[Costantino 99] J. P. Costantino, M. H. Gail, D. Pee, S. Anderson, C. K. Redmond, J. Benichou & H. S. Wieand. *Validation studies for models projecting the risk of invasive and total breast cancer incidence.* Journal of the National Cancer Institute, vol. 91, no. 18, pages 1541–8, 1999. 35, 43, 144

[Cottet 09] V. Cottet, M. Touvier, A. Fournier, M. S. Touillaud, L. Lafay, F. Clavel-Chapelon & M.-C. Boutron-Ruault. *Postmenopausal Breast Cancer Risk and Dietary Patterns in the E3N-EPIC Prospective Cohort Study.* American Journal of Epidemiology, vol. 170, no. 10, pages 1257–1267, 2009. 19

[Couch 97] F. J. Couch, M. L. DeShano, M. A. Blackwood, K. Calzone, J. Stopfer, L. Campeau, A. Ganguly, T. Rebbeck, B. L. Weber, L. Jablon, M. A. Cobleigh, K. Hoskins & J. E. Garber. *BRCA1 Mutations in Women Attending Clinics That Evaluate the Risk*

of Breast Cancer. New England Journal of Medicine, vol. 336, no. 20, pages 1409–1415, 1997. 31

[Cover 67] T. Cover & P. Hart. *Nearest neighbor pattern classification*. Information Theory, IEEE Transactions on, vol. 13, no. 1, pages 21–27, 1967. 111

[D'Agostino 08] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro & W. B. Kannel. *General Cardiovascular Risk Profile for Use in Primary Care*. Circulation, vol. 117, no. 6, pages 743–753, 2008. 27, 28

[Dauplat 04] M.-M. Dauplat & F. Penault-Llorca. *Classification of preinvasive breast and carcinoma in situ : doubts, controversies, and proposal for new categorizations*. Bulletin du Cancer, vol. 91 Suppl 4, pages S205–10, 2004. 16

[De Pauw 09] A. De Pauw, D. Stoppa-Lyonnet, N. Andrieu & B. Asselain. *Estimation du risque individuel de cancer du sein : intérêt et limites des modèles de calcul de risque*. Bulletin du Cancer, vol. 96, no. 10, pages 979–988, 2009. 32

[Decarli 06] Adriano Decarli, Stefano Calza, Giovanna Masala, Claudia Specchia, Domenico Palli & Mitchell H. Gail. *Gail Model for Prediction of Absolute Risk of Invasive Breast Cancer : Independent Evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition Cohort*. Journal of the National Cancer Institute, vol. 98, no. 23, pages 1686–1693, 2006. 36

[Delen 05] Dursun Delen, Glenn Walker & Amit Kadam. *Predicting breast cancer survivability : a comparison of three data mining methods*. Artificial Intelligence in Medicine, vol. 34, no. 2, pages 113–127, 2005. 38, 39, 40

[Dréau 01] H. Dréau, I. Colombet, P. Degoulet & G. Chatellier. *Identification of Patients at High Cardiovascular Risk : a Critical Appraisal of Applicability of Statistical Risk Prediction Models*. Methods of Information in Medicine, vol. 40, pages 6–11, 2001. 29

[Driver 07] J. A. Driver, J. M. Gaziano, R. P. Gelber, I.-M. Lee, J. E. Buring & T. Kurth. *Development of a Risk Score for Colorectal Cancer in Men*. The American Journal of Medicine, vol. 120, no. 3, pages 257–263, 2007. 30

[Dudani 76] S. A. Dudani. *The Distance-Weighted k-Nearest-Neighbor Rule*. Systems, Man and Cybernetics, IEEE Transactions on, vol. SMC-6, no. 4, pages 325–327, 1976. 118

[Dupont 87] W. D. Dupont & D. L. Page. *Breast cancer risk associated with proliferative disease, age at first birth, and a family history of breast cancer.* American Journal of Epidemiology, vol. 125, no. 5, pages 769–779, 1987. 20

[Egan 75] J. P. Egan. Signal detection theory and ROC analysis. Series in Cognition and Perception. Academic Press, 1975. 50

[Endo 08] A. Endo, T. Shibata & H. Tanaka. *Comparison of Seven Algorithms to Predict Breast Cancer Survival.* Biomedical Soft Computing and Human Sciences, vol. 13, no. 2, pages 11–16, 2008. xi, 39

[Eppler 08] M. J. Eppler & M. Aeschimann. *A systematic framework for risk visualization in risk management and communication.* Risk Management (Bas), vol. 11, no. 2, pages 67–89, 2008. xiii, 43, 45, 46

[Fagerström 78] K.O. Fagerström. *Measuring degree of physical dependence to tobacco smoking with reference to individualization of treatment.* Addictive behaviors, vol. 3, no. 3-4, pages 235–241, 1978. 24

[Fagherazzi 11] G. Fagherazzi. *Facteurs alimentaires, composantes du syndrome métabolique et risques de cancer du sein et de diabète de type II dans la cohorte E3N.* PhD thesis, Université Paris XI, 2011. 21

[Fawcett 06] T. Fawcett. *An introduction to ROC analysis.* Pattern Recognition Letters, vol. 27, no. 8, pages 861–874, 2006. xi, 50, 51

[Ferlay 10] J Ferlay, Hr Shin, F Bray, D Forman, C Mathers & DM Parkin. *GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide : IARC CancerBase No. 10*, 2010. xiii, 12, 13, 14

[Fix 51] E. Fix & J. L. Hodges. *Discriminatory analysis, nonparametric discrimination : consistency properties.* Rapport technique, USAF Scholl of aviation and medicine, Randolph Field, 1951. 111, 113

[Flach 03] P. Flach & S. WU. *Repairing concavities in ROC curves.* In Proc. 2003 UK Workshop on Computational Intelligence, pages 38–44, 2003. 51

[Folstein 75] M. F. Folstein, S. E. Folstein & P. R. McHugh. *"Mini-mental state". A practical method for grading the cognitive state of patients for the clinician.* Journal of Psychiatric Research, vol. 12, no. 3, pages 189–198, 1975. 24

[Ford 98] D. Ford, D.F. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D.T. Bishop, B. Weber, G. Lenoir, J. Chang-Claude, H. Sobol, M.D. Teare, J. Struewing, A. Arason, S. Scherneck, J. Peto, T.R. Rebbeck, P. Tonin, S. Neuhausen, R. Barkardottir, J. Eyfjord, H. Lynch, B.A.J. Ponder, S.A. Gayther, J.M. Birch, A. Lindblom, D. Stoppa-Lyonnet, Y. Bignon, A. Borg, U. Hamann, N. Haites, R.J. Scott, C.M. Maugard, H. Vasen, S. Seitz, L.A. Cannon-Albright, A. Schofield & M. Zelada-Hedman. *Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families*. The American Journal of Human Genetics, vol. 62, no. 3, pages 676–689, 1998. 18

[Fournier 05] A. Fournier, F. Berrino, E. Riboli, V. Avenel & F. Clavel-Chapelon. *Breast cancer risk in relation to different types of hormone replacement therapy in the E3N-EPIC cohort*. International Journal of Cancer, vol. 114, no. 3, pages 448–454, 2005. 19, 127

[Fournier 08] A. Fournier, F. Berrino & F. Clavel-Chapelon. *Unequal risks for breast cancer associated with different hormone replacement therapies : results from the E3N cohort study*. Breast Cancer Research and Treatment, vol. 107, no. 1, pages 103–11, 2008. 19

[Friedenreich 08] C. M. Friedenreich & A. E. Cust. *Physical activity and breast cancer risk : impact of timing, type and dose of activity and population subgroup effects*. British Journal of Sports Medicine, vol. 42, no. 8, pages 636–647, 2008. 20

[Gail 89] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer & J. J. Mulvihill. *Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually*. Journal of the National Cancer Institute, vol. 81, no. 24, pages 1879–1886, 1989. xi, 33, 34, 35, 41, 43, 82, 116, 144, 145

[Gandini 00] S. Gandini, H. Merzenich, C. Robertson & P. Boyle. *Meta-analysis of studies on breast cancer risk and diet : the role of fruit and vegetable consumption and the intake of associated micronutrients*. European Journal of Cancer, vol. 36, no. 5, pages 636–646, 2000. 20

[García 08] V. García, R. A. Mollineda & J. S. Sánchez. *A New Performance Evaluation Method for Two-Class Imbalanced Problems*. In Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, SSPR & SPR '08, pages 917–925. Springer-Verlag, 2008. 68

[Gauthier 11] E. Gauthier, L. Brisson, P. Lenca & S. Ragusa. *Breast cancer risk score : a data mining approach to improve readability.* In The International Conference on Data Mining, pages 15–21. CSREA Press, 2011. 132

[Gauthier 12] E. Gauthier, L. Brisson, P. Lenca, F. Clavel-Chapelon & S. Ragusa. *Challenges to building a platform for a breast cancer risk score.* In Sixth International Conference on Research Challenges in Information Science, pages 1–10. IEEE, 2012. 138

[Group 96] Coll. Group. *Breast cancer and hormonal contraceptives : collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Collaborative Group on Hormonal Factors in Breast Cancer.* Lancet, vol. 347, no. 9017, pages 1713–1727, 1996. 19, 115

[Group 97] Coll. Group. *Breast cancer and hormone replacement therapy : collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Collaborative Group on Hormonal Factors in Breast Cancer.* Lancet, vol. 350, no. 9084, pages 1047–1059, 1997. 19, 115

[Group 02] Coll. Group. *Alcohol, tobacco and breast cancer–collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease.* Lancet, vol. 87, no. 11, pages 1234–1245, 2002. 19, 115

[Guessous 10] I. Guessous & S. Durieux-Paillard. *Validation des scores cliniques : notions théoriques et pratiques de base.* Revue Médicale Suisse, no. 264, 2010. xiii, 48, 49

[Hamilton 60] Max Hamilton. *A rating scale for depression.* Journal of Neurology, Neurosurgery & Psychiatry, vol. 23, no. 1, pages 56–62, 1960. 24

[Hanley 82] J. A. Hanley & B. J. McNeil. *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, vol. 143, no. 1, pages 29–36, 1982. 53

[Haq 99] I. U. Haq, L. E. Ramsay, W. W. Yeo, P. R. Jackson & E. J. Wallis. *Is the Framingham risk function valid for northern European populations ? A comparison of methods for estimating absolute coronary risk in high risk men.* Heart, vol. 81, no. 1, pages 40–46, 1999. 29

[Harvard School 08] Public Health (USA) Harvard School. *Disease Risk Index.* Site internet, 2008. xiii, 44

BIBLIOGRAPHIE

[IGAS 03] Inspection générale des affaires sociales IGAS. Santé, pour une politique de prévention durable : rapport annuel 2003. La documentation française, 2003. 7

[Imbs 94] P. Imbs. Trésor de la langue française : dictionnaire de la langue du XIXe et du XXe siècle (1789-1960). Trésor de la langue française : dictionnaire de la langue du XIXe et du XXe siècle. Centre National de la Recherche Scientifique, 1994. 5

[INCa 09] Institut national du cancer INCa. Plan cancer 2009-2013. INCa, Institut national du cancer, 2009. 7

[INSERM 09] Institut National de la Santé et de la Recherche médicale INSERM. Santé des enfants et des adolescents - propositions pour la préserver. Éditions INSERM, 2009. 6

[Jemal 11] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward & D. Forman. *Global cancer statistics*. A Cancer Journal for Clinicians, vol. 61, no. 2, pages 69–90, 2011. 12

[Jerez 05] J. M. Jerez, L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, B. Munarriz & M. Martin. *Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks*. Breast Cancer Research and Treatment, vol. 94, no. 3, pages 265–272, 2005. 40

[Johns 59] M. V. Johns & Columbia Univ. New York Teachers Coll. An empirical bayes approach to non-parametric two-way classification. Defense Technical Information Center, 1959. 111

[Kannel 61] W. B. Kannel, T.R. Dawber, A. Kagan, N. Revotskie & J.I. Stokes. *Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study*. Annals of Internal Medicine, vol. 55, pages 33–50, 1961. 25

[Kannel 67] W. B. Kannel. *Habitual level of physical activity and risk of coronary heart disease : the Framingham study*. Canadian Medical Association Journal, vol. 96, pages 811–812, 1967. 25

[Kannel 76a] W. B. Kannel, M. C. Hjortland, P. M. McNamara & T. Gordon. *Menopause and Risk of Cardiovascular DiseaseThe Framingham Study*. Annals of Internal Medicine, vol. 85, no. 4, pages 447–452, 1976. 25

[Kannel 76b] W. B. Kannel, D. McGee & T. Gordon. *A general cardiovascular risk profile : The Framingham study*. The American Journal of Cardiology, vol. 38, no. 1, pages 46–51, 1976. 25

[Kapoor 05] A. Kapoor & V. G. Vogel. *Prognostic factors for breast cancer and their use in the clinical setting*. Expert Review of Anticancer Therapy, vol. 5, no. 2, pages 269–281, 2005. 16

[Kernbaum 08] S. Kernbaum. Dictionnaire de médecine flammarion. Flammarion, Paris, 8e édition, 2008. 15

[Kesse 02] E. Kesse. *Étude de la relation entre certains facteurs alimentaires et le risque de tumeurs colorectales*. PhD thesis, Institut national agronomique Paris-Grignon, 2002. 68

[Key 88] T.J.A. Key & M.C. Pike. *The role of oestrogens and progestagens in the epidemiology and prevention of breast cancer*. European Journal of Cancer and Clinical Oncology, vol. 24, no. 1, pages 29–43, 1988. 19

[Key 01] T. J. Key, P. K Verkasalo & E. Banks. *Epidemiology of breast cancer*. The Lancet Oncology, vol. 2, no. 3, pages 133–140, 2001. 19, 20

[Kim 06] H.-J. Kim, X. Cui, S. G. Hilsenbeck & A. V. Lee. *Progesterone Receptor Loss Correlates with Human Epidermal Growth Factor Receptor 2 Overexpression in Estrogen Receptor–Positive Breast Cancer*. Clinical Cancer Research, vol. 12, no. 3, pages 1013s–1018s, 2006. 16

[Kivipelto 06] M. Kivipelto, T. Ngandu, T. Laatikainen, B. Winblad, H. Soininen & J. Tuomilehto. *Risk score for the prediction of dementia risk in 20 years among middle aged people : a longitudinal, population-based study*. The Lancet Neurology, vol. 5, no. 9, pages 735–741, 2006. 30

[Kolodner 84] J. L. Kolodner. Retrieval and organizational strategies in conceptual memory : a computer model. Artificial intelligence series. L. Erlbaum Associates, 1984. 110

[Lagrue 02] G. Lagrue, P. Légéron, G. Azoulaï, S. Pelissolo, H. J. Aubin & R. Humbert. *Elaboration d'un test permettant d'évaluer la motivation à l'arrêt du tabac*. Alcoologie et Addictologie, vol. 24, no. 1, pages 33–37, 2002. 24

[Landis 77] J.R. Landis & Koch G.G. *The measurement of observer agreement for categorical data*. Biometrics, vol. 33, no. 1, pages 159–74, 1977. 57

[Laurier 94] D. Laurier, N. Phong Chau, B. Cazelles, P. Segond & the PCV-METRA Group. *Estimation of CHD risk in a french working population using a modified Framingham Model*. Journal of Clinical Epidemiology, vol. 47, no. 12, pages 1353–1364, 1994. 29

[LCC 12] Ligue contre le cancer LCC. *Tout savoir sur le cancer du sein*. Page web, 2012. xiii, 15

[Li 00] C. I. Li, B. O. Anderson, P. Porter, S. K. Holt, J. R. Daling & R. E. Moe. *Changing incidence rate of invasive lobular breast*

carcinoma among older women. Cancer, vol. 88, no. 11, pages 2561–2569, 2000. 19

[Lichtenstein 00] Paul Lichtenstein, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe & Kari Hemminki. *Environmental and Heritable Factors in the Causation of Cancer, Analyses of Cohorts of Twins from Sweden, Denmark, and Finland*. New England Journal of Medicine, vol. 343, no. 2, pages 78–85, 2000. 139

[Longépé 01] C. Longépé. Projet d'urbanisation du système d'information – démarche pratique avec cas concret. Dunod/01 Informatique, 2001. 86

[Luporsi 07] Élisabeth Luporsi & Line Leichtnam-Dugarin. *Comprendre le cancer du sein*. Guide d'information de l'Institut National du Cancer, 2007. 14

[Lynch 11] B. Lynch, C. Friedenreich, E. Winkler, G. Healy, J. Vallance, E. Eakin & N. Owen. *Associations of objectively assessed physical activity and sedentary time with biomarkers of breast cancer risk in postmenopausal women : findings from NHANES (2003–2006)*. Breast Cancer Research and Treatment, vol. 130, pages 183–194, 2011. 20

[MacMahon 80] B. MacMahon, A.P. Andersen, J. Brown, P. Cole, V. Dewaard, T. Kauraniemi, B. Ravhinar, N. Stormby, D. Trichopoulos & K Westlund. *Urine estrogen profiles in European countries with high or low breast cancer rates*. European Journal of Cancer, vol. 16, pages 1627–1632, 1980. 20

[Matsuno 11] R. K. Matsuno, J. P. Costantino, R. G. Ziegler, G. L. Anderson, H. Li, D. Pee & M. H. Gail. *Projecting Individualized Absolute Invasive Breast Cancer Risk in Asian and Pacific Islander American Women*. Journal of the National Cancer Institute, vol. 103, no. 12, pages 951–961, 2011. 35, 43

[Fondation ARC 12] Fondation ARC. La nouvelle fondation ARC affiche son ambition : guérir 2 cancers sur 3 en 2025 ! Communiqué de presse, 2012. 10

[McCormack 06] V. A. McCormack & I. dos Santos Silva. *Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk : A Meta-analysis*. Cancer Epidemiology Biomarkers & Prevention, vol. 15, no. 6, pages 1159–1169, 2006. 19, 133

[Minerva 08] Minerva. *Glossaire des termes utilisés en Evidence-Based Medicine*, 2008. 155

[NCI 11] National Cancer Institute NCI. *Breast Cancer Risk Assessment Tool*. Site internet, 2011. xiii, 44

[Pharoah 97] P. D. P. Pharoah, N. E. Day, S. Duffy, D. F. Easton & B. A. J. Ponder. *Family history and the risk of breast cancer : A systematic review and meta-analysis.* International Journal of Cancer, vol. 71, no. 5, pages 800–809, 1997. 18

[Preston-Martin 90] S. Preston-Martin, M. C. Pike, R. K. Ross, P. A. Jones & B. E. Henderson. *Increased Cell Division as a Cause of Human Cancer.* Cancer Research, vol. 50, no. 23, pages 7415–7421, 1990. 16

[Quinlan 86] J. R. Quinlan. *Induction of decision trees.* Machine Learning, pages 81–106, 1986. 39

[Quinlan 93] Ross Quinlan. C4.5 : Programs for machine learning. Morgan Kaufmann Publishers, 1993. 39

[Remontet 03] L. Remontet, J. Estève, A.M. Bouvier, P. Grosclaude, G. Launoy, F. Menegoz, C. Exbrayat, B. Tretare, P.M. Carli, A.V. Guizard, X. Troussard, P. Bercelli, M. Colonna, J.M. Halna, G. Hedelin, J. Mace-Lesec'h, J. Peng, A. Buemi, M. Velten, E. Jougla, P. Arveux, L. Le Bodic, E. Michel, M. Sauvage, C. Schvartz & J. Faivre. *Cancer incidence and mortality in France over the period 1978-2000.* Revue d'Épidemiologie et de Santé Publique, vol. 51, pages 3–30, 2003. 14

[Reston 03] Va Reston, editeur. Breast imaging reporting and data system atlas (BI-RADS atlas). American College of Radiology, 2003. 36, 65, 123

[Reynolds 04] P. Reynolds, S. Hurley, D. E. Goldberg, H. Anton-Culver, L. Bernstein, D. Deapen, P. L. Horn-Ross, D. Peel, R. Pinder, R. K. Ross, D. West, W. E. Wright & A. Ziogas. *Active Smoking, Household Passive Smoking, and Breast Cancer : Evidence From the California Teachers Study.* Journal of the National Cancer Institute, vol. 96, no. 1, pages 29–37, 2004. 20

[Ringa 08] V. Ringa & A. Fournier. *La diminution de l'utilisation du traitement hormonal de la ménopause a-t-elle fait baisser l'incidence du cancer du sein en France (et ailleurs) ?* Revue d'Épidémiologie et de Santé Publique, vol. 56, no. 5, pages 297–301, 2008. 13

[Rockhill 01] Beverly Rockhill, Donna Spiegelman, Celia Byrne, David J. Hunter & Graham A. Colditz. *Validation of the Gail et al. Model of Breast Cancer Risk Prediction and Implications for Chemoprevention.* Journal of the National Cancer Institute, vol. 93, no. 5, pages 358–366, 2001. 36, 43, 82, 144

BIBLIOGRAPHIE

[Rosner 94] B. Rosner, G. A. Colditz & W. C. Willett. *Reproductive risk factors in a prospective study of breast cancer : the Nurses' Health Study.* American Journal of Epidemiology, vol. 139, no. 8, pages 819–835, 1994. 19

[Rossouw 02] J.E. Rossouw, G.L. Anderson, R.L. Prentice, A.Z. LaCroix, C. Kooperberg, M.L. Stefanick, R.D. Jackson, S.A. Beresford, B.V. Howard, K.C. Johnson, J.M. Kotchen & J. Ockene. *Risks and benefits of estrogen plus progestin in healthy postmenopausal women : Principal results from the women's health initiative randomized controlled trial.* The Journal of the American Medical Association, vol. 288, no. 3, pages 321–333, 2002. 19

[Ruckert 10] F. Ruckert, N. Samm, A.-K. Lehner, H.-D. Saeger, R. Grutzmann & C. Pilarsky. *Simultaneous gene silencing of Bcl-2, XIAP and Survivin re-sensitizes pancreatic cancer cells towards apoptosis.* BioMedCentral Cancer, vol. 10, no. 1, page 379, 2010. xiii, 46

[Russo 00] J. Russo, Y.-F. Hu, X. Yang & I. H. Russo. *Developmental, Cellular, and Molecular Basis of Human Breast Cancer.* Journal of the National Cancer Institute Monographs, vol. 2000, no. 27, pages 17–37, 2000. 16

[Sackett 96] D. L. Sackett, W. M. Rosenberg, J. A. Gray, R. B. Haynes & W. S. Richardson. *Evidence based medicine : what it is and what it isn't.* British Medical Journal, vol. 312, no. 7023, pages 71–72, 1996. 8

[Santen 07] R. J. Santen, N. F. Boyd, R. T. Chlebowski, S. Cummings, J. Cuzick, M. Dowsett, D. Easton, J. F. Forbes, T. Key, S. E. Hankinson, A. Howell & J. Ingle. *Critical assessment of new risk factors for breast cancer : considerations for development of an improved risk prediction model.* Endocrine-Related Cancer, vol. 14, no. 2, pages 169–187, 2007. 33

[Saporta 06] G. Saporta. Probabilités, analyse des données et statistique. Éditions Technip, 2006. 114

[Sassoon 98] J. Sassoon. Urbanisation des systèmes d'information. Management et informatique. Hermès, 1998. 87

[Saunders 93] J. B. Saunders, O. G. Aasland, T. F. Babor, J. R. De La Fuente & M. Grant. *Development of the Alcohol Use Disorders Identification Test (AUDIT) : WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption–II.* Addiction Abingdon England, vol. 88, no. 6, pages 791–804, 1993. 24

[Schwartz 93] D. Schwartz. Méthodes statistiques à l'usage des médecins et des biologistes. Statistique en biologie et en médecine. Flammarion Médecine-Sciences, 1993. 55

[Sebestyen 62] G. S. Sebestyen. Decision-making processes in pattern recognition. ACM monograph series. Macmillan, 1962. 111

[Shattuck-Eidens 97] D. Shattuck-Eidens, A. Oliphant, M. McClure & et al. *Brca1 sequence analysis in women at high risk for susceptibility mutations : Risk factor analysis and implications for genetic testing.* The Journal of the American Medical Association, vol. 278, no. 15, pages 1242–1250, 1997. 31

[Sickles 05] E. A. Sickles, D. L. Miglioretti, R. Ballard-Barbash, B. M. Geller, J. W. T. Leung, R. D. Rosenberg, R. Smith-Bindman & B. C. Yankaskas. *Performance Benchmarks for Diagnostic Mammography.* Radiology, vol. 235, no. 3, pages 775–790, 2005. 64

[Silberschatz 96] A. Silberschatz & A. Tuzhilin. *What Makes Patterns Interesting in Knowledge Discovery Systems.* Knowledge and Data Engineering, IEEE Transactions on, vol. 8, no. 6, pages 970–974, 1996. 71

[Simonin 09] J. Simonin. *Conception de l'architecture d'un système dirigée par un modèle d'urbanisme fonctionnel.* PhD thesis, Université Rennes 1, 2009. 153, 154, 155, 156, 157

[Simonin 11] J. Simonin, E. Bertin, Y. Le Traon, J.M. Jézéquel & Crespi N. *Analysis and Improvement of the Alignment between Business and Information System for Telecom Services.* IARIA International Journal on Advances in Software, vol. 4, no. 42, pages 117–128, 2011. 86

[Simonin 12] J. Simonin, A. Beugnard & Nédélec R. *Processus de développement de système fondé sur l'alignement de modèles.* Revue des sciences et technologies de l'information, vol. 17, no. 3, pages 119–142, 2012. 86, 103

[Stanford 95] J. L. Stanford, L. J. Herrinton, S. M. Schwartz & N. S. Weiss. *Breast cancer incidence in Asian migrants to the United States and their descendants.* Epidemiology (Cambridge, Mass), vol. 6, no. 2, pages 181–183, 1995. 20

[Teasdale 74] G. Teasdale & B. Jennett. *Assessment of coma and impaired consciousness : a practical scale.* The Lancet, vol. 304, no. 7872, pages 81–84, 1974. 24

[Tehard 02] B. Tehard, M. J. Van Liere, C. C. Nougué & F. Clavel-Chapelon. *Anthropometric measurements and body silhouette of women : validity and perception.* Journal of the American

Dietetic Association, vol. 102, no. 12, pages 1779–1784, 2002. 68

[Tehard 06] B. Tehard, C. M. Friedenreich, J.-M. Oppert & F. Clavel-Chapelon. *Effect of Physical Activity on Women at Increased Risk of Breast Cancer : Results from the E3N Cohort Study.* Cancer Epidemiology Biomarkers & Prevention, vol. 15, no. 1, pages 57–64, 2006. 20

[Tice 05] J. A. Tice, S. R. Cummings, E. Ziv & K. Kerlikowske. *Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population.* Breast Cancer Research and Treatment, vol. 94, no. 2, pages 115–122, 2005. 36

[Tice 08] J. A. Tice, S. R. Cummings, R. Smith-Bindman, L. Ichikawa, W. E. Barlow & K. Kerlikowske. *Using Clinical Factors and Mammographic Breast Density to Estimate Breast Cancer Risk : Development and Validation of a New Predictive Model.* Annals of Internal Medicine, vol. 148, no. 5, pages 337–347, 2008. 36

[Timsit 05] J. F. Timsit, C. Alberti & S. Chevret. *Cox proportional hazards regression analysis.* Revue des Maladies Respiratoires, vol. 22, no. 6 Pt 1, pages 1058–1064, 2005. 27

[Trétarre 04] B. Trétarre, A. V. Guizard, D. Fontaine, membres du réseau FRANCIM & le CépiDc-Inserm. *Cancer du sein chez la femme : incidence et mortalité, France 2000.* Bulletin Épidémiologique Hebdomadaire, vol. 44, pages 209–210, 2004. xiii, 18

[Tyrer 04] Jonathan Tyrer, Stephen W. Duffy & Jack Cuzick. *A breast cancer prediction model incorporating familial and personal risk factors.* Statistics in Medicine, vol. 23, no. 7, pages 1111–1130, 2004. 32

[USC 57] United States Commission USC. *Commission on chronic Illness.* Harvard University Press, Cambridge, Mass., 1957. 5

[Varghese 12] J. S. Varghese, D. J. Thompson, K. Michailidou, S. Lindström, C. Turnbull, J. Brown, J. Leyland, R. M. L. Warren, R. N. Luben, R. J. Loos, N. J. Wareham, J. Rommens, A. D. Paterson, L. J. Martin, C. M. Vachon, C. G. Scott, E. J. Atkinson, F. J. Couch, C. Apicella, M. C. Southey, J. Stone, J. Li, L. Eriksson, K Czene, N. F. Boyd, P. Hall, J. L. Hopper, R. M. Tamimi, N. Rahman & D. F. Easton. *Mammographic Breast Density and Breast Cancer : Evidence of a Shared Genetic Basis.* Cancer Research, vol. 72, no. 6, pages 1478–1484, 2012. 19, 133

[Vercambre 09] M. N. Vercambre, M.-C. Boutron-Ruault, M. Niravong, C. Berr, F. Clavel-Chapelon & S. Ragusa. *Performance of a short dietary questionnaire to assess nutrient intake using regression-based weights.* Public Health Nutrition, vol. 12, no. 4, pages 547–552, 2009. 24

[Vergnaud 08] A.-C. Vergnaud, S. Bertrais, P. Galan, S. Hercberg & S. Czernichow. *Ten-year risk prediction in French men using the Framingham coronary score : Results from the national SU.VI.MAX cohort.* Preventive Medicine, vol. 47, no. 1, pages 61–65, 2008. 30

[Weiss 03] G. M. Weiss & Provost F. J. *Learning when training data are costly : the effect of class distribution on tree induction.* Journal of Artificial Intelligence Research, 2003. 69

[Weiss 04] G. M. Weiss. *Mining with rarity : a unifying framework.* SIGKDD Explorations Newsletter, vol. 6, no. 1, pages 7–19, 2004. 69

[WHO 00] World Health Organization WHO. *Obesity : preventing and managing the global epidemic. Report of a WHO consultation.* World Health Organization Technical Report Series, vol. 894, no. 7, pages i–xii, 1–253, 2000. 23

[Xue 07] F. Xue & K. B. Michels. *Diabetes, metabolic syndrome, and breast cancer : a review of the current evidence.* The American Journal of Clinical Nutrition, vol. 86, no. 3, pages 823S–835S, 2007. 21

Technopôle Brest-Iroise - CS 83818
29238 Brest Cedex 3
France
Tél : + 33 (0)2 29 00 11 11
www.telecom-bretagne.eu