Apports d'une méthode de fouille de données pour la détection des cas incidents de cancer du sein dans les données du Programme de Médicalisation des Systèmes d'Information : une analyse formelle des concepts sur les doonées 2001 de PMSI et du registre du cancer de l'Isère

Christophe Goetz

UNIVERSITÉ HENRI POINCARÉ, NANCY 1 FACULTÉ DE MÉDECINE DE NANCY
2011 N°

THÈSE

pour obtenir le grade de

DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement

dans le cadre du troisième cycle de Médecine Spécialisée

par

Christophe GOETZ

le 19 mai 2011

Apports d'une méthode de fouille de données pour la détection des cas incidents de cancer du sein dans les données du Programme de Médicalisation des Systèmes d'Information :

Une analyse formelle des concepts sur les données 2001 du PMSI et du registre du cancer de l'Isère

Examinateurs de la thèse :

Monsieur F. KOHLER	Professeur	Président
Monsieur F. GUILLEMIN	Professeur	Juges
Monsieur P. CHASTAGNER	Professeur	
Monsieur N. JAY	Maître de Conférences	

UNIVERSITÉ HENRI POINCARÉ, NANCY 1
2011

FACULTÉ DE MÉDECINE DE NANCY
N°

THÈSE

pour obtenir le grade de

DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement

dans le cadre du troisième cycle de Médecine Spécialisée

par

Christophe GOETZ

le 19 mai 2011

Apports d'une méthode de fouille de données pour la détection des cas incidents de cancer du sein dans les données du Programme de Médicalisation des Systèmes d'Information :

Une analyse formelle des concepts sur les données 2001 du PMSI et du registre du cancer de l'Isère

Examinateurs de la thèse :

Monsieur F. KOHLER	Professeur	Président
Monsieur F. GUILLEMIN	Professeur	Juges
Monsieur P. CHASTAGNER	Professeur	
Monsieur N. JAY	Maître de Conférences	

UNIVERSITÉ HENRI POINCARÉ, NANCY 1

FACULTÉ DE MÉDECINE DE NANCY

Président de l'Université : Professeur Jean-Pierre FINANCE

Doyen de la Faculté de Médecine : Professeur Henry COUDANE

Vice Doyen *Mission « sillon lorrain »* : Professeur Annick BARBAUD
Vice Doyen *Mission « Campus »* : Professeur Marie-Christine BÉNÉ
Vice Doyen *Mission « Finances » :* Professeur Marc BRAUN
Vice Doyen Mission « *Recherche* » : Professeur Jean-Louis GUÉANT

Assesseurs :

- Pédagogie :	**Professeur Karine ANGIOÏ-DUPREZ**
- 1er Cycle :	**Professeur Bernard FOLIGUET**
- « Première année commune aux études de santé (PACES) et universitarisation études para-médicales »	**M. Christophe NÉMOS**
- 2ème Cycle :	**Professeur Marc DEBOUVERIE**
- 3ème Cycle :	
« DES Spécialités Médicales, Chirurgicales et Biologiques »	**Professeur Jean-Pierre BRONOWICKI**
« DES Spécialité Médecine Générale	**Professeur Francis RAPHAËL**
- Filières professionnalisées :	**M. Walter BLONDEL**
- Formation Continue :	**Professeur Hervé VESPIGNANI**
- Commission de Prospective :	**Professeur Pierre-Edouard BOLLAERT**
- Recherche :	**Professeur Didier MAINARD**
- Développement Professionnel Continu :	**Professeur Jean-Dominique DE KORWIN**

DOYENS HONORAIRES

Professeur Adrien DUPREZ – Professeur Jean-Bernard DUREUX
Professeur Jacques ROLAND – Professeur Patrick NETTER

==========

PROFESSEURS HONORAIRES

Pierre ALEXANDRE – Jean-Marie ANDRE - Daniel ANTHOINE - Alain BERTRAND - Pierre BEY - Jacques BORRELLY Michel BOULANGE - Jean-Claude BURDIN - Claude BURLET - Daniel BURNEL - Claude CHARDOT Jean-Pierre CRANCE - Gérard DEBRY - Jean-Pierre DELAGOUTTE - Emile de LAVERGNE - Jean-Pierre DESCHAMPS Michel DUC - Jean DUHEILLE - Adrien DUPREZ - Jean-Bernard DUREUX - Gérard FIEVE - Jean FLOQUET - Robert FRISCH Alain GAUCHER - Pierre GAUCHER - Hubert GERARD - Jean-Marie GILGENKRANTZ - Simone GILGENKRANTZ Oliéro GUERCI - Pierre HARTEMANN - Claude HURIET - Christian JANOT - Jacques LACOSTE - Henri LAMBERT Pierre LANDES - Alain LARCAN - Marie-Claire LAXENAIRE - Michel LAXENAIRE - Jacques LECLERE - Pierre LEDERLIN Bernard LEGRAS - Michel MANCIAUX - Jean-Pierre MALLIÉ – Philippe MANGIN - Pierre MATHIEU Denise MONERET-VAUTRIN - Pierre NABET - Jean-Pierre NICOLAS - Pierre PAYSANT - Francis PENIN Gilbert PERCEBOIS Claude PERRIN - Guy PETIET - Luc PICARD - Michel PIERSON - Jean-Marie POLU – Jacques POUREL Jean PREVOT Antoine RASPILLER - Michel RENARD - Jacques ROLAND - René-Jean ROYER - Paul SADOUL - Daniel SCHMITT Michel SCHWEITZER - Jean SOMMELET - Danièle SOMMELET - Michel STRICKER - Gilbert THIBAUT Augusta TREHEUX Hubert UFFHOLTZ - Gérard VAILLANT - Paul VERT - Colette VIDAILHET - Michel VIDAILHET - Michel WAYOFF Michel WEBER

==========

PROFESSEURS DES UNIVERSITÉS
PRATICIENS HOSPITALIERS

(Disciplines du Conseil National des Universités)

42ème Section : MORPHOLOGIE ET MORPHOGENÈSE
1ère sous-section : *(Anatomie)*
Professeur Gilles GROSDIDIER
Professeur Pierre LASCOMBES – Professeur Marc BRAUN
2ème sous-section : *(Cytologie et histologie)*
Professeur Bernard FOLIGUET
3ème sous-section : *(Anatomie et cytologie pathologiques)*
Professeur François PLENAT – Professeur Jean-Michel VIGNAUD

43ème Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE

1ère sous-section : ***(Biophysique et médecine nucléaire)***

Professeur Gilles KARCHER – Professeur Pierre-Yves MARIE – Professeur Pierre OLIVIER

2ème sous-section : ***(Radiologie et imagerie médicale)***

Professeur Denis REGENT – Professeur Michel CLAUDON

Professeur Serge BRACARD – Professeur Alain BLUM – Professeur Jacques FELBLINGER

Professeur René ANXIONNAT

44ème Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION

1ère sous-section : ***(Biochimie et biologie moléculaire)***

Professeur Jean-Louis GUÉANT – Professeur Jean-Luc OLIVIER – Professeur Bernard NAMOUR

2ème sous-section : ***(Physiologie)***

Professeur François MARCHAL – Professeur Bruno CHENUEL – Professeur Christian BEYAERT

3ème sous-section : ***(Biologie Cellulaire)***

Professeur Ali DALLOUL

4ème sous-section : ***(Nutrition)***

Professeur Olivier ZIEGLER – Professeur Didier QUILLIOT

45ème Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE

1ère sous-section : ***(Bactériologie – virologie ; hygiène hospitalière)***

Professeur Alain LE FAOU - Professeur Alain LOZNIEWSKI

3ème sous-section : ***(Maladies infectieuses ; maladies tropicales)***

Professeur Thierry MAY – Professeur Christian RABAUD

46ème Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ

1ère sous-section : ***(Épidémiologie, économie de la santé et prévention)***

Professeur Philippe HARTEMANN – Professeur Serge BRIANÇON - Professeur Francis GUILLEMIN

Professeur Denis ZMIROU-NAVIER – Professeur François ALLA

2ème sous-section : ***(Médecine et santé au travail)***

Professeur Christophe PARIS

3ème sous-section : ***(Médecine légale et droit de la santé)***

Professeur Henry COUDANE

4ème sous-section : ***(Biostatistiques, informatique médicale et technologies de communication)***

Professeur François KOHLER – Professeur Éliane ALBUISSON

47ème Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE

1ère sous-section : ***(Hématologie ; transfusion)***

Professeur Thomas LECOMPTE – Professeur Pierre BORDIGONI

Professeur Jean-François STOLTZ – Professeur Pierre FEUGIER

2ème sous-section : ***(Cancérologie ; radiothérapie)***

Professeur François GUILLEMIN – Professeur Thierry CONROY

Professeur Didier PEIFFERT – Professeur Frédéric MARCHAL

3ème sous-section : ***(Immunologie)***

Professeur Gilbert FAURE – Professeur Marie-Christine BENE

4ème sous-section : ***(Génétique)***

Professeur Philippe JONVEAUX – Professeur Bruno LEHEUP

48ème Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE

1ère sous-section : ***(Anesthésiologie et réanimation chirurgicale ; médecine d'urgence)***

Professeur Claude MEISTELMAN – Professeur Hervé BOUAZIZ

Professeur Paul-Michel MERTES – Professeur Gérard AUDIBERT

2ème sous-section : ***(Réanimation médicale ; médecine d'urgence)***

Professeur Alain GERARD - Professeur Pierre-Édouard BOLLAERT

Professeur Bruno LÉVY – Professeur Sébastien GIBOT

3ème sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique ; addictologie)***

Professeur Patrick NETTER – Professeur Pierre GILLET

4ème sous-section : ***(Thérapeutique ; médecine d'urgence ; addictologie)***

Professeur François PAILLE – Professeur Gérard GAY – Professeur Faiez ZANNAD - Professeur Patrick ROSSIGNOL

49ème Section : PATHOLOGIE NERVEUSE ET MUSCULAIRE, PATHOLOGIE MENTALE, HANDICAP et RÉÉDUCATION

1ère sous-section : ***(Neurologie)***

Professeur Gérard BARROCHE – Professeur Hervé VESPIGNANI
Professeur Xavier DUCROCQ – Professeur Marc DEBOUVERIE

2ème sous-section : ***(Neurochirurgie)***

Professeur Jean-Claude MARCHAL – Professeur Jean AUQUE
Professeur Thierry CIVIT

3ème sous-section : ***(Psychiatrie d'adultes ; addictologie)***

Professeur Jean-Pierre KAHN – Professeur Raymund SCHWAN

4ème sous-section : ***(Pédopsychiatrie ; addictologie)***

Professeur Daniel SIBERTIN-BLANC – Professeur Bernard KABUTH

5ème sous-section : ***(Médecine physique et de réadaptation)***

Professeur Jean PAYSANT

50ème Section : PATHOLOGIE OSTÉO-ARTICULAIRE, DERMATOLOGIE et CHIRURGIE PLASTIQUE

1ère sous-section : ***(Rhumatologie)***

Professeur Isabelle CHARY-VALCKENAERE – Professeur Damien LOEUILLE

2ème sous-section : ***(Chirurgie orthopédique et traumatologique)***

Professeur Daniel MOLE - Professeur Didier MAINARD
Professeur François SIRVEAUX – Professeur Laurent GALOIS

3ème sous-section : ***(Dermato-vénéréologie)***

Professeur Jean-Luc SCHMUTZ – Professeur Annick BARBAUD

4ème sous-section : ***(Chirurgie plastique, reconstructrice et esthétique ; brûlologie)***

Professeur François DAP – Professeur Gilles DAUTEL

51ème Section : PATHOLOGIE CARDIORESPIRATOIRE et VASCULAIRE

1ère sous-section : ***(Pneumologie ; addictologie)***

Professeur Yves MARTINET – Professeur Jean-François CHABOT – Professeur Ari CHAOUAT

2ème sous-section : ***(Cardiologie)***

Professeur Etienne ALIOT – Professeur Yves JUILLIERE – Professeur Nicolas SADOUL
Professeur Christian de CHILLOU

3ème sous-section : ***(Chirurgie thoracique et cardiovasculaire)***

Professeur Jean-Pierre VILLEMOT - Professeur Jean-Pierre CARTEAUX

4ème sous-section : ***(Chirurgie vasculaire ; médecine vasculaire)***

Professeur Denis WAHL – Professeur Sergueï MALIKOV

52ème Section : MALADIES DES APPAREILS DIGESTIF et URINAIRE

1ère sous-section : ***(Gastroentérologie ; hépatologie ; addictologie)***

Professeur Marc-André BIGARD - Professeur Jean-Pierre BRONOWICKI – Professeur Laurent PEYRIN-BIROULET

2ème sous-section : ***(Chirurgie digestive)***

3ème sous-section : ***(Néphrologie)***

Professeur Michèle KESSLER – Professeur Dominique HESTIN – Professeur Luc FRIMAT

4ème sous-section : ***(Urologie)***

Professeur Jacques HUBERT – Professeur Pascal ESCHWEGE

53ème Section : MÉDECINE INTERNE, GÉRIATRIE et CHIRURGIE GÉNÉRALE

1ère sous-section : ***(Médecine interne ; gériatrie et biologie du vieillissement ; médecine générale ; addictologie)***

Professeur Jean-Dominique DE KORWIN – Professeur Pierre KAMINSKY
Professeur Athanase BENETOS - Professeur Gisèle KANNY – Professeur Christine PERRET-GUILLAUME

2ème sous-section : ***(Chirurgie générale)***

Professeur Patrick BOISSEL – Professeur Laurent BRESLER
Professeur Laurent BRUNAUD – Professeur Ahmet AYAV

54ème Section : DÉVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNÉCOLOGIE-OBSTÉTRIQUE, ENDOCRINOLOGIE ET REPRODUCTION

1ère sous-section : ***(Pédiatrie)***

Professeur Pierre MONIN - Professeur Jean-Michel HASCOET - Professeur Pascal CHASTAGNER
Professeur François FEILLET - Professeur Cyril SCHWEITZER

2ème sous-section : ***(Chirurgie infantile)***

Professeur Michel SCHMITT – Professeur Pierre JOURNEAU – Professeur Jean-Louis LEMELLE

3ème sous-section : ***(Gynécologie-obstétrique ; gynécologie médicale)***

Professeur Jean-Louis BOUTROY - Professeur Philippe JUDLIN – Professeur Patricia BARBARINO

4ème sous-section : ***(Endocrinologie, diabète et maladies métaboliques ; gynécologie médicale)***

Professeur Georges WERYHA – Professeur Marc KLEIN – Professeur Bruno GUERCI

55ème Section : PATHOLOGIE DE LA TÊTE ET DU COU

1ère sous-section : ***(Oto-rhino-laryngologie)***

Professeur Claude SIMON – Professeur Roger JANKOWSKI – Professeur Cécile PARIETTI-WINKLER

2ème sous-section : ***(Ophtalmologie)***

Professeur Jean-Luc GEORGE – Professeur Jean-Paul BERROD – Professeur Karine ANGIOI-DUPREZ

3ème sous-section : ***(Chirurgie maxillo-faciale et stomatologie)***

Professeur Jean-François CHASSAGNE – Professeur Etienne SIMON

==========

PROFESSEURS DES UNIVERSITÉS

64ème Section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE

Professeur Sandrine BOSCHI-MULLER

==========

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

42ème Section : MORPHOLOGIE ET MORPHOGENÈSE

1ère sous-section : ***(Anatomie)***

Docteur Bruno GRIGNON – Docteur Thierry HAUMONT – Docteur Manuela PEREZ

2ème sous-section : ***(Cytologie et histologie)***

Docteur Edouard BARRAT - Docteur Françoise TOUATI – Docteur Chantal KOHLER

3ème sous-section : ***(Anatomie et cytologie pathologiques)***

Docteur Aude BRESSENOT

43ème Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE

1ère sous-section : ***(Biophysique et médecine nucléaire)***

Docteur Marie-Hélène LAURENS – Docteur Jean-Claude MAYER
Docteur Pierre THOUVENOT – Docteur Jean-Marie ESCANYE

2ème sous-section : (Radiologie et imagerie médicale)

Docteur Damien MANDRY

44ème Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION

1ère sous-section : ***(Biochimie et biologie moléculaire)***

Docteur Jean STRACZEK – Docteur Sophie FREMONT
Docteur Isabelle GASTIN – Docteur Marc MERTEN – Docteur Catherine MALAPLATE-ARMAND
Docteur Shyue-Fang BATTAGLIA

3ème sous-section : ***(Biologie Cellulaire)***

Docteur Véronique DECOT-MAILLERET

4ème sous-section : ***(Nutrition)***

Docteur Rosa-Maria RODRIGUEZ-GUEANT

45ème Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE
1ère sous-section : ***(Bactériologie – Virologie ; hygiène hospitalière)***
Docteur Francine MORY – Docteur Véronique VENARD
2ème sous-section : ***(Parasitologie et mycologie)***
Docteur Nelly CONTET-AUDONNEAU – Madame Marie MACHOUART

46ème Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ
1ère sous-section : ***(Epidémiologie, économie de la santé et prévention)***
Docteur Alexis HAUTEMANIÈRE – Docteur Frédérique CLAUDOT
3ème sous-section ***(Médecine légale et droit de la santé)***
Docteur Laurent MARTRILLE
4ère sous-section : ***(Biostatistiques, informatique médicale et technologies de communication)***
Docteur Nicolas JAY

47ème Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE
2ème sous-section : ***(Cancérologie ; radiothérapie : cancérologie (type mixte : biologique)***
Docteur Lina BOLOTINE
3ème sous-section : ***(Immunologie)***
Docteur Marcelo DE CARVALHO BITTENCOURT
4ème sous-section : ***(Génétique)***
Docteur Christophe PHILIPPE – Docteur Céline BONNET

48ème Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE
3ème sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique)***
Docteur Françoise LAPICQUE – Docteur Marie-José ROYER-MORROT – Docteur Nicolas GAMBIER

50ème Section : RHUMATOLOGIE
1ère sous-section : ***(Rhumatologie)***
Docteur Anne-Christine RAT
3ème sous-section : ***(Dermato-vénéréologie)***
Docteur Anne-Claire BURSZTEJN

54ème Section : DÉVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNÉCOLOGIE-OBSTÉTRIQUE, ENDOCRINOLOGIE ET REPRODUCTION
3ème sous-section :
Docteur Olivier MOREL
5ème sous-section : ***(Biologie et médecine du développement et de la reproduction ; gynécologie médicale)***
Docteur Jean-Louis CORDONNIER

==========

MAÎTRES DE CONFÉRENCES

5ème section : SCIENCE ÉCONOMIE GÉNÉRALE
Monsieur Vincent LHUILLIER

40ème section : SCIENCES DU MÉDICAMENT
Monsieur Jean-François COLLIN

60ème section : MÉCANIQUE, GÉNIE MÉCANIQUE ET GÉNIE CIVILE
Monsieur Alain DURAND

61ème section : GÉNIE INFORMATIQUE, AUTOMATIQUE ET TRAITEMENT DU SIGNAL
Monsieur Jean REBSTOCK – Monsieur Walter BLONDEL

64ème section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
Mademoiselle Marie-Claire LANHERS – Monsieur Pascal REBOUL – Mr Nick RAMALANJAONA

65ème section : BIOLOGIE CELLULAIRE
Mademoiselle Françoise DREYFUSS – Monsieur Jean-Louis GELLY
Madame Ketsia HESS – Monsieur Hervé MEMBRE – Monsieur Christophe NEMOS - Madame Natalia DE ISLA
Madame Nathalie MERCIER

66ème section : PHYSIOLOGIE
Monsieur Nguyen TRAN

67ème section : BIOLOGIE DES POPULATIONS ET ÉCOLOGIE
Madame Nadine MUSSE

==========

PROFESSEURS ASSOCIÉS

Médecine Générale
Professeur associé Alain AUBREGE
Professeur associé Francis RAPHAEL

MAÎTRES DE CONFÉRENCES ASSOCIÉS

Médecine Générale
Docteur Jean-Marc BOIVIN
Docteur Jean-Louis ADAM
Docteur Elisabeth STEYER
Docteur Paolo DI PATRIZIO
Docteur Sophie SIEGRIST

==========

PROFESSEURS ÉMÉRITES

Professeur Jean-Marie ANDRÉ - Professeur Daniel ANTHOINE - Professeur Pierre BEY - Professeur Michel BOULANGÉ
Professeur Jean-Pierre CRANCE – Professeur Jean-Pierre DELAGOUTTE - Professeur Jean-Marie GILGENKRANTZ
Professeur Simone GILGENKRANTZ - Professeur Henri LAMBERT - Professeur Alain LARCAN
Professeur Denise MONERET-VAUTRIN - Professeur Jean-Pierre NICOLAS - Professeur Luc PICARD
Professeur Michel PIERSON - Professeur Jacques POUREL - Professeur Jacques ROLAND – Professeur Michel STRICKER
Professeur Gilbert THIBAUT - Professeur Hubert UFFHOLTZ - Professeur Paul VERT - Professeur Colette VIDAILHET
Professeur Michel VIDAILHET

==========

DOCTEURS HONORIS CAUSA

Professeur Norman SHUMWAY (1972)
Université de Stanford, Californie (U.S.A)
Professeur Paul MICHIELSEN (1979)
Université Catholique, Louvain (Belgique)
Professeur Charles A. BERRY (1982)
Centre de Médecine Préventive, Houston (U.S.A)
Professeur Pierre-Marie GALETTI (1982)
Brown University, Providence (U.S.A)
Professeur Mamish Nisbet MUNRO (1982)
Massachusetts Institute of Technology (U.S.A)
Professeur Mildred T. STAHLMAN (1982)
Vanderbilt University, Nashville (U.S.A)
Harry J. BUNCKE (1989)
Université de Californie, San Francisco (U.S.A)
Professeur Daniel G. BICHET (2001)
Université de Montréal (Canada)
Professeur Brian BURCHELL (2007)
Université de Dundee (Royaume Uni)

Professeur Théodore H. SCHIEBLER (1989)
Institut d'Anatomie de Würtzburg (R.F.A)
Professeur Maria DELIVORIA-PAPADOPOULOS (1996)
Université de Pennsylvanie (U.S.A)
Professeur Mashaki KASHIWARA (1996)
Research Institute for Mathematical Sciences de Kyoto (JAPON)
Professeur Ralph GRÄSBECK (1996)
Université d'Helsinki (FINLANDE)
Professeur James STEICHEN (1997)
Université d'Indianapolis (U.S.A)
Professeur Duong Quang TRUNG (1997)
Centre Universitaire de Formation et de Perfectionnement des Professionnels de Santé d'Hô Chi Minh-Ville (VIÊTNAM)

Professeur Marc LEVENSTON (2005)
Institute of Technology, Atlanta (USA)

REMERCIEMENTS

À notre maître et président de thèse,

Monsieur le professeur François KOHLER,
Professeur de biostatistiques, informatique médicale et technologies de communication.

Vous m'avez fait l'honneur de présider ce jury.
J'ai été honoré de bénéficier de vos enseignements et de votre expertise de l'information médicale tout au long de mes études médicales.
Soyez assuré de ma plus grande gratitude.

À notre maître et juge,

Monsieur le professeur Francis GUILLEMIN,

Professeur d'épidémiologie, économie de la santé et prévention.

Vous m'avez fait l'honneur de juger cette thèse.
Vous m'avez guidé au cours de mon internat sur les voies de l'épidémiologie et de la psychométrie.
Veuillez trouver ici l'expression de mon plus grand respect.

À notre maître et juge,

Monsieur le professeur Pascal CHASTAGNER,

Professeur de pédiatrie, service d'onco-hématologie pédiatrique.

Vous m'avez fait l'honneur de juger cette thèse.
Votre expertise clinique et scientifique pour juger de ce travail m'est précieuse.
Recevez ici le témoignage de ma profonde gratitude.

À notre directeur et juge,

Monsieur le Docteur Nicolas JAY,

Maître de conférences de biostatistiques, informatique médicale et technologies de communication.

Tu m'as fais l'honneur de me confier de sujet, qui est une occasion formidable de travailler à la croisée de deux domaines qui me tiennent à cœur : l'épidémiologie et l'information médicale.
Trouve ici l'expression de ma profonde reconnaissance.

À Aurélien ZANG, pour sa précieuse collaboration à ce travail et sa maitrise des logiciels d'analyse formelle de concepts.

Au groupe ONC-EPI et au CNADS, pour leur éclairage et la mise à notre disposition des données utilisées, et pour le financement de notre participation au congrès EMOIS 2011.

À toutes les équipes des services qui m'ont accueilli au cours de mon internat, qui m'ont toujours accompagné avec attention et beaucoup appris.

À la formidable équipe du DIMS du CHR de Metz-Thionville, pour leur accueil chaleureux et le travail fait ensemble durant cette dernière année.

À tous les internes de santé publique de Nancy et de l'inter-région Nord-est, pour le chemin parcouru ensemble ces quatre dernières années.

À Emilie, pour son soutien infaillible et indispensable de chaque instant.

À mon grand père, dont la gentillesse et la simplicité sont une source de motivation intarissable.

À mes parents, pour tout.

À mes beaux parents, pour leur fantastique dynamisme et leur accueil plus que chaleureux.

À mon filleul, Pierre, pour sa bonne humeur si rafraichissante.

À mon frère, pour son intarissable soif de sommets.

À toute ma famille,

À mes amis, Damien, Delphine, Laurent, Manue, Marion, Olivier, Romain, Vincent, et tous les autres, parce qu'avec eux la vie est toujours un peu plus colorée.

SERMENT

« Au moment d'être admis à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité.

Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux.

Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité.

J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences. Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences.

Je donnerai mes soins à l'indigent et à quiconque me le demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçu à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs.

Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré et méprisé si j'y manque. »

TABLE DES MATIERES

Il est très triste qu'il y ait si peu d'informations inutiles de nos jours.

[Oscar Wilde]

CHAPITRE I

–

Le Programme de Médicalisation des Systèmes d'Information

Le Programme de Médicalisation des Systèmes d'Information (PMSI) trouve sa définition dans la loi du 31 juillet 1991 portant réforme hospitalière [1] : il s'agit pour les établissements de santé français, publics comme privés, de pouvoir transmettre aux autorités de tutelle (services de l'État et Assurance Maladie) des analyses de leur activité qui tiennent compte des aspects médicaux de celle-ci (pathologies et modes de prise en charge). Il s'agit d'une évolution par rapport aux indicateurs utilisés jusqu'à lors (nombre de lits, nombre de journées, etc.) qui n'intégraient aucune information de nature médicale.

L'objectif du PMSI est double. Il facilite d'une part la planification sanitaire et est notamment utilisé pour le suivi des schémas régionaux d'organisation sanitaire (SROS). Il permet d'autre part d'améliorer la maitrise des dépenses de santé. Ainsi depuis 2004, le PMSI s'inscrit dans le cadre de la Tarification À l'Activité (T2A) [2] : les données d'activité fournies par les établissements sont directement utilisées pour calculer les ressources qui leur seront allouées par l'Assurance Maladie.

Dans le champ de court séjour de Médecine, Chirurgie, Obstétrique et odontologie (MCO), le fonctionnement du PMSI a été inspiré des Diagnosis Related Groups (DRG) définis par Robert B. Fetter au début des années 1980 pour le système de santé américain [3]. Le recueil est organisé au niveau des unités médicales : à l'issue du passage d'un malade dans une unité médicale, un Résumé d'Unité Médicale (RUM) est produit. Il contient les informations administratives et médicales suivantes :

- le numéro administratif local de séjour,
- le sexe du patient, sa date de naissance et le code postal de son lieu de résidence,
- le code et le type d'autorisation de l'unité médicale
- le numéro de l'établissement dans le Fichier national des établissements sanitaires et sociaux (Finess),
- les dates et modes d'entrée et de sortie, provenance et destination,
- le nombre de séances,
- les diagnostics codés selon la 10ème révision de la Classification Internationale des Maladies (CIM 10) :
    - le diagnostic principal est celui qui a occasionné l'effort de soin le plus important pour l'unité médicale[1],
    - le diagnostic relié sert à préciser le diagnostic principal lorsque celui-ci le nécessite,
    - les diagnostics associés significatifs décrivent l'ensemble des autres problèmes de santé qui ont été pris en charge dans l'unité médicale,
- les actes médicaux réalisés, codés selon la Classification Commune des Actes Médicaux (CCAM), ou le Catalogue des Actes Médicaux (CdAM) avant 2005,
- le poids à l'entrée dans l'unité médicale pour les nouveau-nés,
- l'âge gestationnel pour la mère et le nouveau-né,
- l'indice de gravité simplifié (IGS II),
- les données associées documentaires.

[1] Depuis mars 2010, le diagnostic principal est le problème de santé qui a motivé l'admission dans l'unité médicale.

Le séjour hospitalier d'un patient est ainsi constitué d'un ou plusieurs RUM (si différentes unités médicales sont fréquentées), qui sont « groupés » ensemble par un programme officiel fourni aux établissements, la fonction de groupage, pour produire un Résumé Standardisé de Sortie (RSS). La fonction de groupage classe ensuite le RSS dans un Groupe Homogène de Malades (GHM). Le classement des RSS en GHM est exhaustif est unique : tout RSS est obligatoirement classé dans un GHM et dans un seul. Il repose sur des critères médicaux (diagnostics et actes principalement) mais aussi économiques : un GHM regroupe des séjours dont les consommations en ressources sont voisines. Pour transmettre les données du PMSI aux autorités de tutelle, les RSS sont rendus anonymes sous la forme de Résumé de Sortie Anonyme (RSA) : les dates d'entrée et de sortie sont remplacées par la durée du séjour et le mois de sortie, la date de naissance est transformée en âge et le code postal est remplacé par un code géographique. Depuis 2001, un numéro de chaînage anonyme est ajouté afin de pouvoir relier entre eux les différents séjours d'un même patient [4].

Ainsi le PMSI constitue une formidable base de données médico-administrative sur les séjours hospitaliers, dont la vocation première est certes financière, mais dont la richesse et la couverture exhaustive de l'ensemble du territoire français semblent pouvoir servir d'autres intérêts, en premier lieu épidémiologiques. Les biais potentiels de cette utilisation du PMSI à des fins d'épidémiologie sont nombreux et bien connus [5,6] : confusion entre cas incidents et cas prévalents, erreurs de codage, non exhaustivité du recueil des comorbidités, non congruence des modèles économiques et médicaux… Mais la littérature florissante sur le sujet tend à prouver qu'ils sont tout sauf insurmontables [7-11].

CHAPITRE II
–
Épidémiologie du cancer du sein

Le cancer du sein représente depuis de nombreuses années un problème de santé publique dans la plupart des pays développés [12,13]. Avec 50 000 nouveaux cas en 2005 en France, il reste le cancer le plus fréquent chez la femme [14]. Son incidence a presque doublé depuis 1980 (figure 1), en partie en raison l'amélioration de sa détection grâce à la mise en place de campagnes d'information et surtout d'un programme de dépistage national proposant une mammographie tous les deux ans aux femmes de 50 à 74 ans [15].

Durant les 20 dernières années, les prises en charges précoces et les progrès thérapeutiques [13] (amélioration de la chirurgie et de la radiothérapie, développement de la chimiothérapie et de l'hormonothérapie) ont permis de commencer à faire reculer la mortalité du cancer du sein (figure 1), même si on lui attribuait encore plus de 11 000 décès en 2005 en France [14]. Mais si ces progrès ont pu être réalisés, c'est également grâce à des données de surveillance épidémiologiques fiables, précises et réactives, qui ont permis de guider des politiques et des programmes de santé publique adaptés et efficaces.

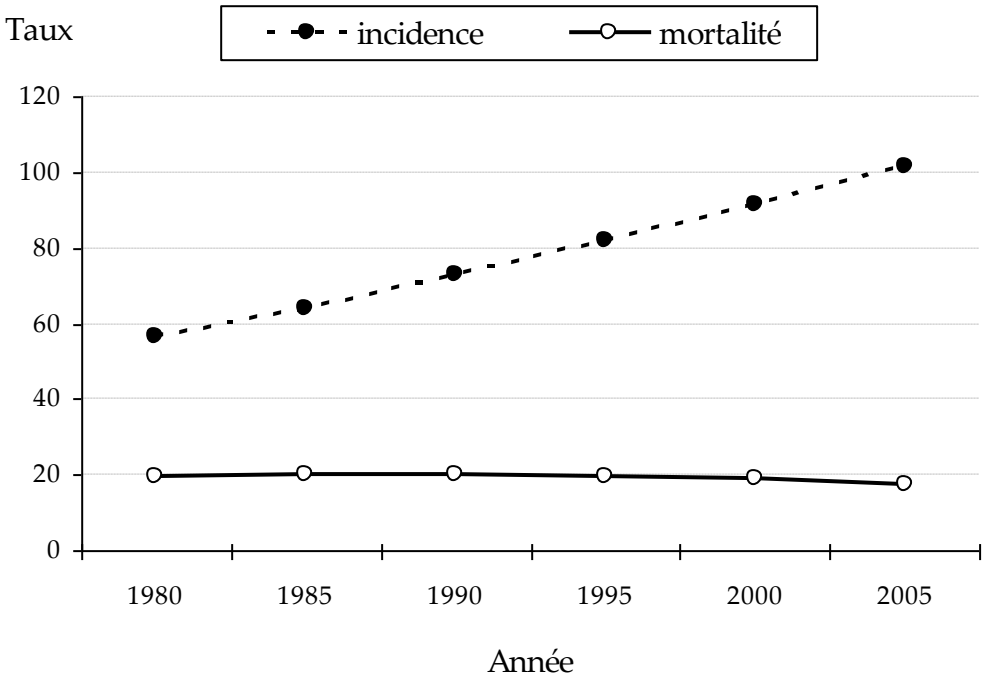

Figure 1 : Évolution des taux d'incidence et de mortalité du cancer du sein en France de 1980 à 2005 (standardisés monde pour 100 000 personnes-années). Données InVS – Francim – CépiDc – HCL [14].

La surveillance épidémiologique du cancer du sein comme des autres cancers repose sur les données du réseau français des registres du cancer (Francim) et celles du Centre d'épidémiologie sur les causes médicales de décès de l'Institut national de la santé et de la recherche médicale (CépiDc-Inserm). Les registres réalisent l'enregistrement exhaustif des nouveaux cas de cancer grâce notamment aux données des laboratoires d'anatomopathologie, des réseaux de cancérologie, des régimes d'assurance maladie et du PMSI. Ces cas sont ensuite confirmés et documentés grâce aux dossiers médicaux et aux états civils des mairies.

Si les données du CépiDc couvrent la France entière, ce n'est pas le cas des registres du cancer qui se limitent à quelques exceptions près[2] à une zone géographique limitée, en général un département. Ainsi pour le cancer du sein en

[2] Les hémopathies et les tumeurs solides de l'enfant font l'objet de registres nationaux : Registre National des Hémopathies Malignes de l'Enfant (RNHE) et Registre National des Tumeurs Solides de l'Enfant (RNTSE).

2011, 14 registres généraux et un registre spécialisé permettent de couvrir seulement 16 des 96 départements métropolitains, soit environ 20% de la population française. L'incidence au niveau national est une estimation calculée d'après le rapport entre l'incidence et la mortalité mesuré dans les départements couverts par un registre, la mortalité étant fournie au niveau national par le CépiDc. Cette estimation suppose que les pratiques diagnostiques et la survie pour le cancer du sein sont comparables entre les différents départements, ce qui est fortement discutable [16].

Une autre limite des registres est une latence importante avant la disponibilité des données. En effet, pour chaque cas de cancer signalé, un retour au dossier médical est nécessaire avant l'inclusion afin de confirmer le diagnostic et de le documenter le cas échéant. Cette procédure lourde, mais qui assure une excellente qualité des données, entraine un délai de plusieurs années avant que celles-ci ne soient exploitables. Ainsi en 2011, les données disponibles les plus récentes étaient celles de l'année 2005.

Ainsi, les registres du cancer sont la référence pour étudier l'incidence du cancer du sein, mais ils présentent deux limites : ils ne couvrent pas toute le territoire français et leurs données sont disponibles avec une latence de plusieurs années. Le PMSI, qui est déjà une source d'information incontournable pour alimenter les registres, semble pouvoir pallier ces limites car il est exhaustif sur toute la France et ses données sont disponibles d'une année sur l'autre.

CHAPITRE III

–

Utilisation de bases de données médico-administratives pour estimer l'incidence du cancer du sein

Depuis les années 1980, la volonté de maîtriser les dépenses de santé a fait naître des bases de données médico-administratives dans de multiples pays, telles le PMSI en France ou Medicare[3] aux États-Unis. L'utilisation de ces bases de données, réputées exhaustives et facilement disponibles, pour la recherche de cas incidents de cancer a fait l'objet de nombreux travaux de recherche [9, 10, 16, 17-46]. Le cancer du sein a été particulièrement étudié, notamment car il relève très souvent d'une prise en charge hospitalière avec un parcours de soin relativement bien identifié [10, 16, 31-46].

Ce qui ressort de la littérature est la difficulté d'obtenir grâce aux bases de données médico-administratives des estimations de l'incidence aussi précises et aussi fiables que celles issues des registres du cancer. Dans le cas du cancer du sein, les études mettent en évidence des sensibilités allant de 75 à 90% et des spécificités allant de 99,75 à 99,86% [35]. Les sources d'erreur sont multiples. La première est la confusion entre les cas incidents et les cas prévalents : il est parfois difficile de dire si une hospitalisation donnée correspond à un nouveau cancer ou à la suite de la prise en charge d'un cancer déjà connu (cancers récidivants, cancers synchrones, etc.). Le codage de l'information dans les bases de données médico-administratives est également en cause : les défaillances en termes d'exhaustivité et de qualité du codage augmentent le risque d'erreur, même si la situation pour le PMSI s'est améliorée depuis 2007 (date à partir de laquelle la part de la T2A dans le financement des séjours hospitaliers est passée à 100%). Les parcours de soins

[3] Medicare est un système d'assurance santé pour les plus de 65 ans géré par le gouvernement des États-Unis qui produit une base de données utilisant la classification des DRG [3].

atypiques sont une autre source d'erreur. En effet, certains bilans diagnostics de cancer du sein sont réalisés en médecine de ville, hors de l'hôpital. Certains cancers ne sont pas opérés. D'autres ne sont même jamais hospitalisés : il est alors impossible de les retrouver dans le PMSI. Les estimations d'incidences sont également biaisées par la différence possible entre la date réelle de diagnostic du cancer (inconnue des bases médico-administratives) et la date de la première hospitalisation pour ce cancer (utilisée par substitution). Les cas ne sont alors pas toujours attribués à la bonne période d'étude.

Pour améliorer les estimations d'incidence issues des données des bases médico-administratives, deux approches sont possibles : effectuer des corrections a posteriori sur les estimations obtenues afin d'en corriger les biais, ou réduire les erreurs à la source en améliorant les qualités des algorithmes d'identification des cas incidents.

De nombreuses méthodes de correction de l'incidence ont été proposées [16, 31-32, 42-46], mais la construction des algorithmes d'identification des cas est souvent une étape négligée car elle relève de choix empiriques réalisés d'après les pratiques de prise en charge attendues et les règles de codage des bases médico-administratives concernées.

Les deux algorithmes les plus couramment retrouvés pour l'identification des cas incidents de cancer du sein sont la recherche d'un diagnostic principal de cancer du sein, et la recherche de l'association diagnostic principal de cancer du sein et acte de mastectomie [10, 16, 35-37, 42-45]. Un curage ganglionnaire est parfois également recherché [37]. Pour éliminer davantage de cas prévalents, certains algorithmes prévoient d'exclure les patientes déjà présentes dans la base de données avec un diagnostic de cancer du sein les années précédentes [37, 41] mais cela suppose d'avoir un chainage efficace sur une durée suffisante, ce qui est souvent une difficulté importante [11].

Certains algorithmes ont fait l'objet de constructions plus élaborées. On peut par exemple citer Freeman & al [31] qui ont proposé un algorithme d'après les résultats d'une régression logistique sur 4 combinaisons de 36 informations (codes de diagnostics et de procédures médicales) avec les données de Medicare, en comparaison à un registre de cancer. Mais la sélection arbitraire de 36 codes parmi les dizaines de milliers possibles reste une limite importante. Nattinger & al [32] ont proposé un algorithme très complet, en 4 étapes successives permettant d'identifier des cas incidents potentiels de cancer du sein, puis de les confirmer ou de les infirmer en fonction d'informations complémentaires ; mais là encore l'algorithme repose sur une sélection *a priori* d'un nombre limité d'information d'intérêt.

La difficulté pour la construction de ces algorithmes est la nécessité de composer avec d'une part les aléas des prises en charges thérapeutiques des patients et d'autre part les aléas du codage de ces prises en charges dans les bases médico-administratives, qui utilise des classifications de diagnostics et d'actes médicaux permettant un nombre exponentiel de combinaisons possibles. Pour s'affranchir de l'étape empirique dans la construction des algorithmes, il faut donc pouvoir traiter l'ensemble de ces combinaisons. Cela est impossible avec les méthodes statistiques usuelles telle que la régression logistique, mais il existe des méthodes spécifiquement adaptées à cette problématique : celles de l'Extraction de Connaissances à partir de Données.

CHAPITRE IV
–
L'Extraction de Connaissances à partir de Données

Durant les dernières décennies, l'informatisation a – dans le domaine de la santé comme ailleurs – facilité la génération, la manipulation et le stockage d'un grand nombre d'informations. En découle un volume de données toujours plus important, qui a entrainé avec lui une adaptation des outils statistiques nécessaires à leur exploitation. Parmi eux, l'Extraction de Connaissances à partir de Données (ECD) est déjà largement répandue dans les domaines de l'industrie et des finances et tend à se populariser dans le domaine de la santé publique [47-49].

L'objectif de l'ECD est d'identifier, dans des volumes importants de données, des relations jusqu'à lors inconnues, « cachées », pour aboutir à une connaissance nouvelle [50]. Cette approche se distingue donc fondamentalement des statistiques dites « classiques », qui consistent en la vérification d'hypothèses formulées *a priori*. L'ECD fonctionne selon un processus cyclique dans lequel on distingue généralement 5 étapes :

1. la sélection de données cible,
2. la préparation des données en vue du traitement (nettoyage et prétraitement),
3. la réduction et la projection des données (choix et transformation des variables),
4. l'utilisation de techniques de fouille de données (ou « data mining ») pour identifier des motifs d'intérêt,
5. l'évaluation et la validation des résultats.

Les nouvelles connaissances issues de l'étape 5 peuvent éventuellement être utilisées pour démarrer un nouveau cycle, entier ou partiel (figure 2).

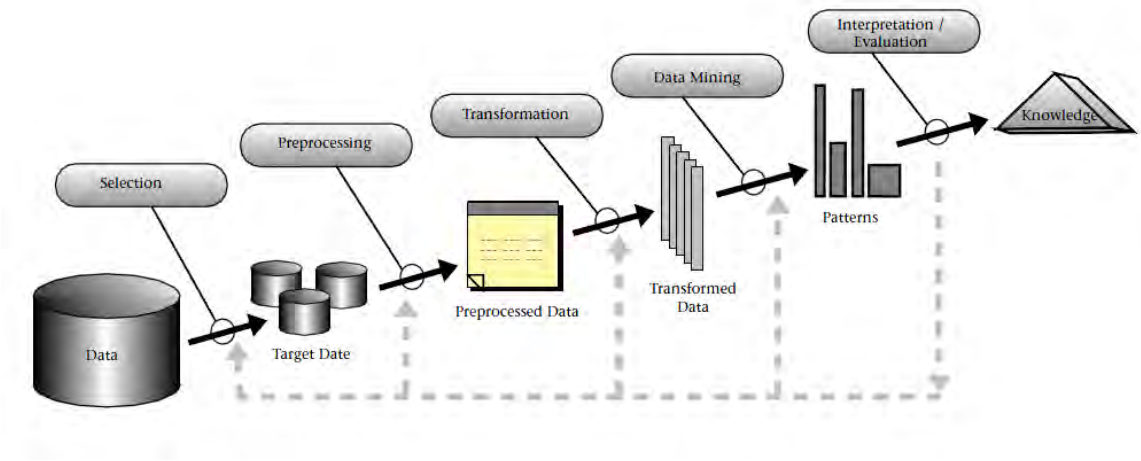

Figure 2 : Le processus d'extraction de connaissances à partir de données, issu de [51].

Parmi les méthodes de fouille de données, l'Analyse Formelle de Concepts (AFC) est une méthode permettant de représenter des données de façon hiérarchique. Introduite en 1982 par Rudolf Wille [52] sous la forme d'une application de la théorie des treillis, l'AFC est également connue sous la dénomination de « treillis de Galois ».

Pour réaliser une AFC, il est nécessaire de définir un contexte formel. En mathématique un contexte définit des objets, des attributs et leurs relations. On peut notamment le représenter sous la forme d'un tableau à double entrée présentant les objets en lignes et les attributs en colonne (figure 3) ; chaque case permet alors de décrire de façon binaire la relation en un objet donné et un attribut donné.

	Cancer du sein	Métastases	Chimiothérapie	Chirurgie	Radiothérapie
Patiente A	×	×	×		
Patiente B	×			×	×
Patiente C	×	×	×	×	
Patiente D	×		×		
Patiente E	×			×	×

Figure 3 : Contexte formel décrivant les relations entre 5 objets : les patientes A, B, C, D et E, et 5 attributs : cancer du sein, métastases, chimiothérapie, chirurgie et radiothérapie.

L'AFC est une méthode qui permet de détecter automatiquement des regroupements naturels d'un ensemble d'objets liés à un ensemble d'attributs. On appelle ces regroupements des concepts formels. Un concept formel est ainsi un couple d'ensembles d'objets et d'attributs (A, B), tel que les attributs B sont communs aux objets A, et réciproquement les objets A possèdent les attributs B. Par exemple selon le contexte formel présenté dans la figure 3, on peut décrire le concept formel ({Patiente A, Patiente C}, {Cancer du sein, Métastases, Chimiothérapie}) : les patientes A et C ont exactement en commun les attributs (cancer du sein, métastases, chimiothérapie) ; aucune autre patiente ne partage tous ces attributs et aucun autre attribut n'est partagé par ces patientes. On appelle intension l'ensemble des attributs du concept et extension l'ensemble des objets du concept. Par exemple le concept ({Patiente A, Patiente C}, {Cancer du sein, Métastases, Chimiothérapie}) a une intension contenant 3 attributs (Cancer du sein, Métastases, Chimiothérapie) et une extension contenant 2 objets (Patiente A, Patiente B).

Après avoir identifié les concepts formels, l'AFC permet d'en construire une hiérarchie, sous la forme d'un treillis, selon la relation d'ordre qui existe entre eux (figure 4). Ainsi, un concept C1 est inférieur à un concept C2 si l'extension de C1 est comprise dans l'extension de C2. Le résultat est un treillis partiellement ordonné car tous les concepts ne sont pas directement comparables entre eux. Par

exemple dans le treillis présenté en figure 4, le concept ({Patiente A, Patiente C}, {Cancer du sein, Métastases, Chimiothérapie}) et le concept ({Patiente B, Patiente E}, {Cancer du sein, Chirurgie, Radiothérapie}) ne peuvent pas être directement comparés.

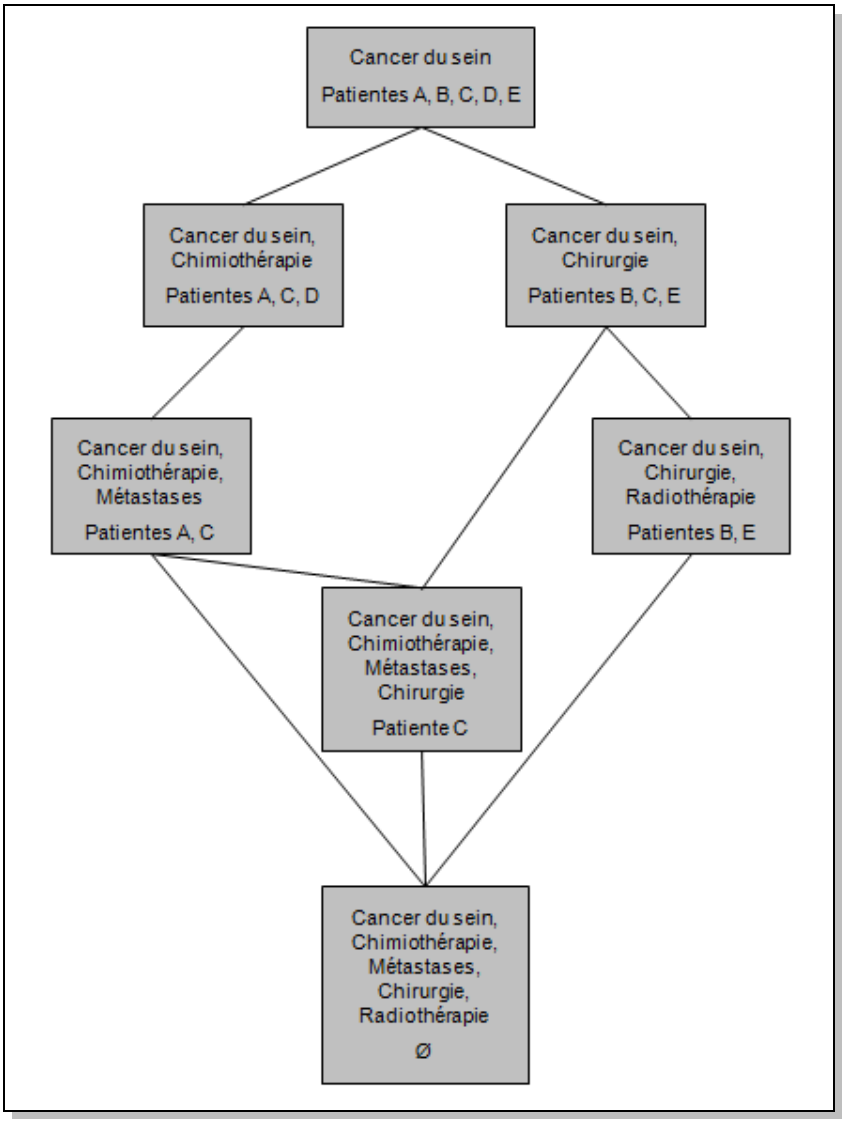

Figure 4 : Treillis des concepts issus du contexte formel (figure 3).

La construction de hiérarchies de concepts à partir d'un ensemble d'observations et leur visualisation sous la forme d'un treillis font ainsi de l'AFC une méthode de fouille de données idéale, tout particulièrement pour

l'exploitation de bases de données contenant de nombreuses variables qualitatives. Avec ses multiples variables sociodémographiques et médicales, dont des diagnostics codés selon les plus de 17 000 de possibilités offertes par la CIM 10 et des actes selon les plus de 7 600 codes de la CCAM, l'exploitation épidémiologique du PMSI – et en particulier pour la détection de cas incidents de cancer du sein – semble avoir tout à gagner des méthodes d'ECD telles que l'AFC.

CHAPITRE V
–
Apports d'une méthode de fouille de données pour la détection des cancers du sein incidents dans les données du Programme de Médicalisation des Systèmes d'Information

Article accepté le 12/12/2010 pour publication dans la revue *Informatique et Santé.*

Christophe Goetz[1], Aurélien Zang[1], le groupe ONC-EPI[2] et Nicolas Jay[1,3]

[1]*Laboratoire SPI-EAO, Faculté de médecine de Nancy, Vandœuvre-lès-Nancy, France*
[2]*Observatoire des cancers en région Rhône Alpes*
[3]*Orpailleur, LORIA, Vandœuvre-lès-Nancy, France*

Introduction

Le cancer du sein est un enjeu de santé publique majeur dans la plupart des pays développés. Durant la dernière décennie, d'importants progrès ont été réalisés dans sa prévention et dans sa prise en charge [12]. La poursuite de cette dynamique nécessite un suivi régulier des données épidémiologiques, tant pour organiser les filières de soins que pour guider la mise en place de nouvelles actions et en apprécier l'efficacité.

Cette surveillance épidémiologique est idéalement réalisée au travers de registres populationnels de cancer [53]. Les registres fournissent des données fiables mais le processus de validation des cas de cancer inclus requiert du temps et il existe donc une latence de plusieurs années avant que les données ne soient exploitables. De plus, leur organisation étant complexe et onéreuse, les registres ne couvrent souvent que des territoires géographiques restreints. Il n'existe ainsi pas de registre national pour le cancer du sein en France, ni dans de nombreux pays

développés (États-Unis, Royaume-Uni, Irlande, Espagne, Italie et Suisse entre autres). En France en 2008, 11 registres généraux et un registre spécialisé permettaient de couvrir seulement 14 des 96 départements métropolitains.

Contrairement aux registres, les bases de données médico-administratives sont des sources d'information exhaustives sur de larges territoires et fournissent des données rapidement exploitables. Elles ont été adoptées dans de nombreux pays et ont pour objectif de quantifier l'activité médicale, sur la base des Diagnosis Related Groups (DRG) [3]. En France, c'est le Programme de Médicalisation des Systèmes d'Information (PMSI) qui regroupe les informations médico-administratives de l'ensemble des séjours hospitaliers du pays. Mais les difficultés posées par l'utilisation de ce type de données pour l'estimation de l'incidence sont multiples et déjà bien connues [5, 54]. La principale concerne la confusion entre cas incidents et cas prévalents, mais on peut également citer les biais dus à l'objectif économique des données (tarification de l'activité), les erreurs de codage et l'existence de très nombreux codes pour décrire les diagnostics et les actes.

De nombreux travaux ont déjà été réalisés pour produire des indicateurs épidémiologiques du cancer du sein grâce aux bases de données médico-administratives [35]. Diverses méthodes ont été proposées pour estimer et corriger les biais [16, 31, 32, 42-46], mais tous les critères d'identification utilisés (qu'on appellera « algorithmes » dans cet article) partent d'une sélection a priori de codes de diagnostics ou d'actes. Le très grand nombre de codes de diagnostics et d'actes utilisés dans les bases médico-administratives empêche en effet d'exploiter toute l'information disponible avec des méthodes statistiques classiques. Un nouvel éclairage et peut-être de nouvelles solutions pourraient être apportés par les méthodes d'Extraction de Connaissances à partir de Données (ECD), qui permettent une exploitation automatique ou semi-automatique de bases de données volumineuses [50]. Ces méthodes ont déjà montré leur utilité pour l'exploitation des données médico-administratives, notamment en

pharmacovigilance [47,48] et pour l'analyse de parcours de soins [49]. Elles pourraient être utilisées pour mettre au point un algorithme optimal d'identification des cas incidents de cancer du sein dans le PMSI.

Notre objectif était donc d'évaluer les apports d'une méthode d'ECD, l'Analyse Formelle de Concepts (AFC), en tant qu'outil de classification épidémiologique des données du PMSI. Ainsi, notre étude porte sur la découverte semi-automatisée des attributs du PMSI marqueurs de cas incidents de cancer du sein.

Matériel et méthodes

Données

Les données utilisées provenaient du PMSI et du registre du cancer de l'Isère pour l'année 2001. Elles ont déjà fait l'objet d'une publication en tant que base de validation pour une étude sur l'incidence du cancer du sein [44].

Les données du PMSI utilisées pour cette étude comprenaient l'ensemble des Résumés d'Unité Médicale (RUM) des femmes de 20 ans et plus s'achevant entre le 1er janvier et le 31 décembre 2001, avec un code postal de résidence de l'Isère (38) et un code diagnostique de cancer invasif du sein quelle que soit sa position (principal, relié ou associé) dans tous les établissements de France métropolitaine.

Les données du registre du cancer de l'Isère utilisées pour cette étude correspondaient à l'ensemble des femmes de 20 ans et plus ayant eu un diagnostic de cancer invasif du sein en 2001 et résidant en Isère au moment du diagnostic.

Les données du PMSI ont été confrontées à celles du registre grâce au logiciel Anonymat [55]. Au total, le registre a identifié 825 cas incidents de cancer du sein et le PMSI 1639 patientes hospitalisées avec un diagnostic de cancer du sein (figure 5).

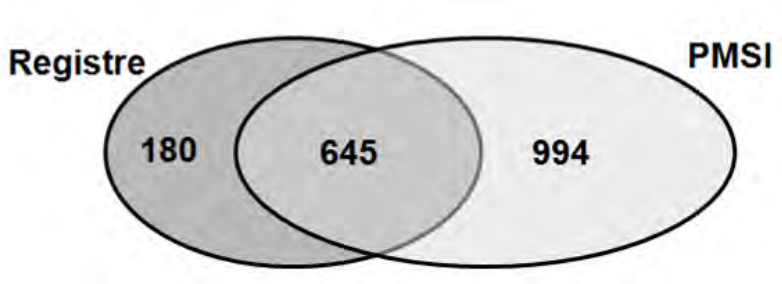

Figure 5 : Cas distincts et cas communs entre les cancers du sein incidents présents dans le registre et les hospitalisations pour cancer du sein dans le PMSI, femmes de 20 ans et plus, année 2001, Isère.

Analyses

L'extraction de connaissances à partir de données (ECD) consiste en un processus cyclique : sélection de données, prétraitement, transformation, fouille des données, visualisation et interprétation des résultats, qui peuvent donner lieu à un nouveau cycle [50].

L'analyse formelle de concepts (AFC) [52] est une méthode de classification non supervisée qui permet de traiter les données d'un contexte formel décrivant des associations binaires entre des objets et leurs attributs (figure 6A). L'AFC organise ces objets et ces attributs en regroupements naturels appelés concepts formels. Un concept est couple formé d'un ensemble d'objets, dénommé extension, et d'un ensemble d'attributs, dénommé intension. Pour un concept donné, tous les objets de l'intension ont en commun tous les attributs de l'extension. L'ensemble des concepts associés à un contexte formel peut être ordonné et représenté dans un treillis de Galois (figure 6B).

A

	cancer sein	chirurgie	radiothérapie	chimiothérapie
patiente1	X		X	
patiente2			X	
patiente3	X	X		X

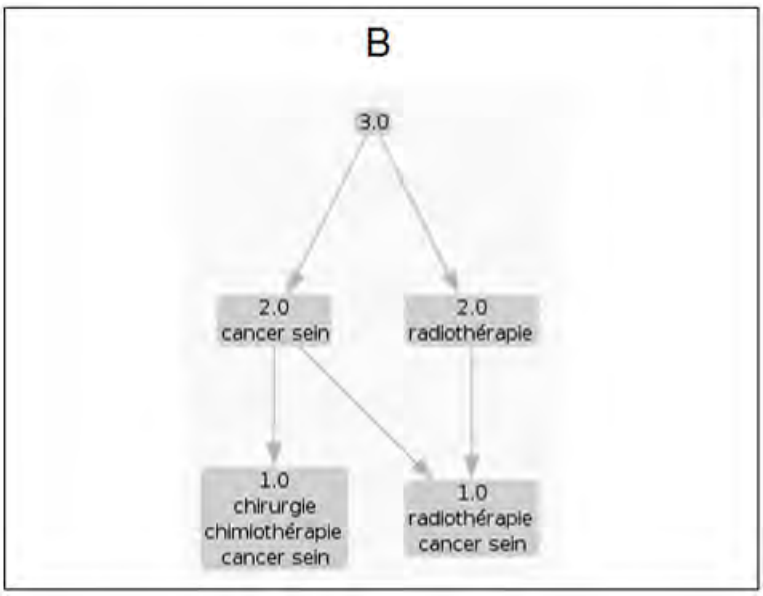

Figure 6 : Exemple de contexte formel (A) avec 3 objets (les patientes) et 4 attributs, et son treillis de Galois (B) présentant 5 concepts et leurs effectifs.

Pour l'analyse des données du PMSI, les objets étaient les patientes et les attributs étaient les informations disponibles dans les RUM de tous leurs séjours de l'année 2001. Les attributs retenus pour les analyses étaient :

- la nature et le type des diagnostics (principal ou non),
- les actes,
- l'âge, découpé en 3 classes selon l'épidémiologie du cancer du sein [56] et selon le programme national de dépistage systématique [13] : moins de 50 ans (cancers précoces), de 50 à 75 ans (période du dépistage systématique), et plus de 75 ans (cancers tardifs),
- le trimestre de sortie du premier séjour,
- les établissements dans lesquels la patiente a séjourné.

L'intension de chacun des concepts formés d'après ces objets et ces attributs a été considérée comme un algorithme candidat pour la sélection des cas incidents de cancer du sein. Ainsi, une intension contenant les attributs A et B correspondrait à un algorithme recherchant les patientes ayant présenté les caractéristiques A ET B dans l'année.

On appelle effectif d'un concept le nombre de patientes dans son extension. Pour chaque concept, les données du registre ont servi de référence pour calculer sensibilité et Valeur Prédictive Positive (VPP). L'incidence du cancer étant très faible (2 cas pour 1000 habitantes de 20 ans et plus en Isère en 2001), la spécificité et la VPN étaient peu pertinentes pour juger la qualité des concepts car toujours proches de 1. La sensibilité représentait la capacité d'un concept à écarter les faux négatifs, et la VPP sa capacité à écarter les faux positifs.

Résultats

L'AFC a permis d'identifier 66 555 concepts. Le treillis de Galois hiérarchisant ces algorithmes met en évidence deux groupes de concepts (figure 7). Le premier groupe (partie droite du treillis) fait ressortir principalement le diagnostic de chimiothérapie, ainsi que les séjours en début d'année (premier et deuxième trimestre). Le second groupe (partie gauche du treillis) met en évidence le diagnostic principal de cancer du sein ainsi que des actes correspondant à une prise en charge chirurgicale pour cancer du sein : mastectomie, curage ganglionnaire, anesthésie et passage en salle de réveil.

Le tableau 1 donne les caractéristiques (VPP et Sensibilité) des concepts d'intérêt.

Les attributs associés aux concepts présentant les meilleures sensibilités étaient le diagnostic principal de cancer du sein, l'acte d'anesthésie générale, le passage en salle de réveil, un âge compris entre 50 et 75 ans, l'acte de curage ganglionnaire axillaire unilatéral et le code diagnostic de chimiothérapie.

Les attributs associés aux concepts présentant les valeurs prédictives les plus hautes étaient le diagnostic principal de cancer du sein, le diagnostic de chimiothérapie, et les actes de curage ganglionnaire, anesthésie, passage en salle de réveil et mastectomie partielle.

Les attributs associés aux concepts présentant les VPP les plus basses étaient le diagnostic principal de chimiothérapie, les diagnostics chimiothérapie et de tumeurs secondaires (métastases), et les séjours avec une date de sortie dans le premier trimestre de l'année.

Avec une sensibilité de 68% et une VPP de 69%, le concept qui présentait le meilleur compromis pour minimiser les faux positifs et les faux négatifs était la présence d'un code de cancer du sein en position de diagnostic principal (figure 8).

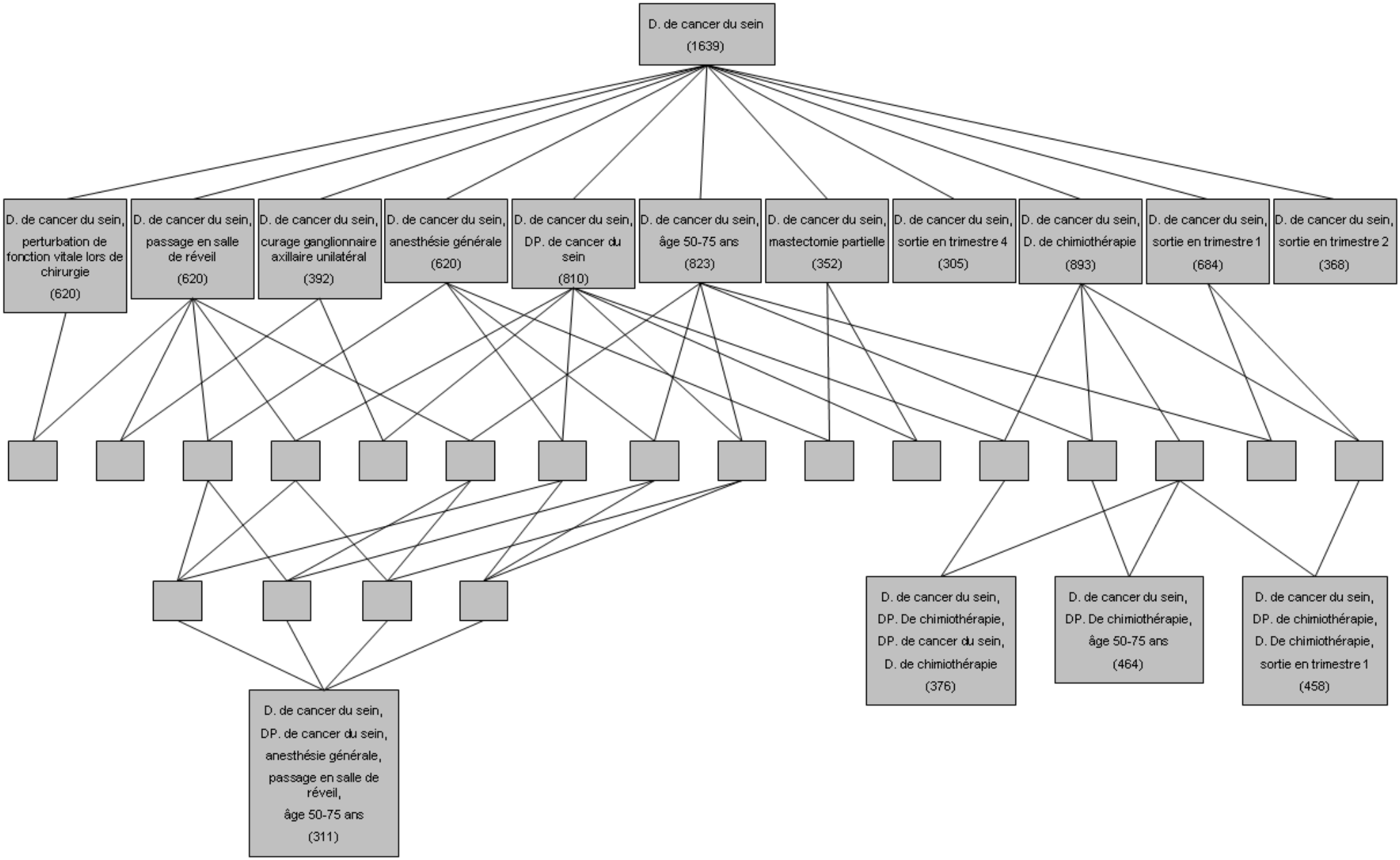

Figure 7 : Treillis de Galois représentant les concepts d'effectif supérieur à 300, construits à partir des données du PMSI (D. = diagnostic, DP. = diagnostic principal). Les concepts intermédiaires ne sont pas explicités pour des raisons de lisibilité.

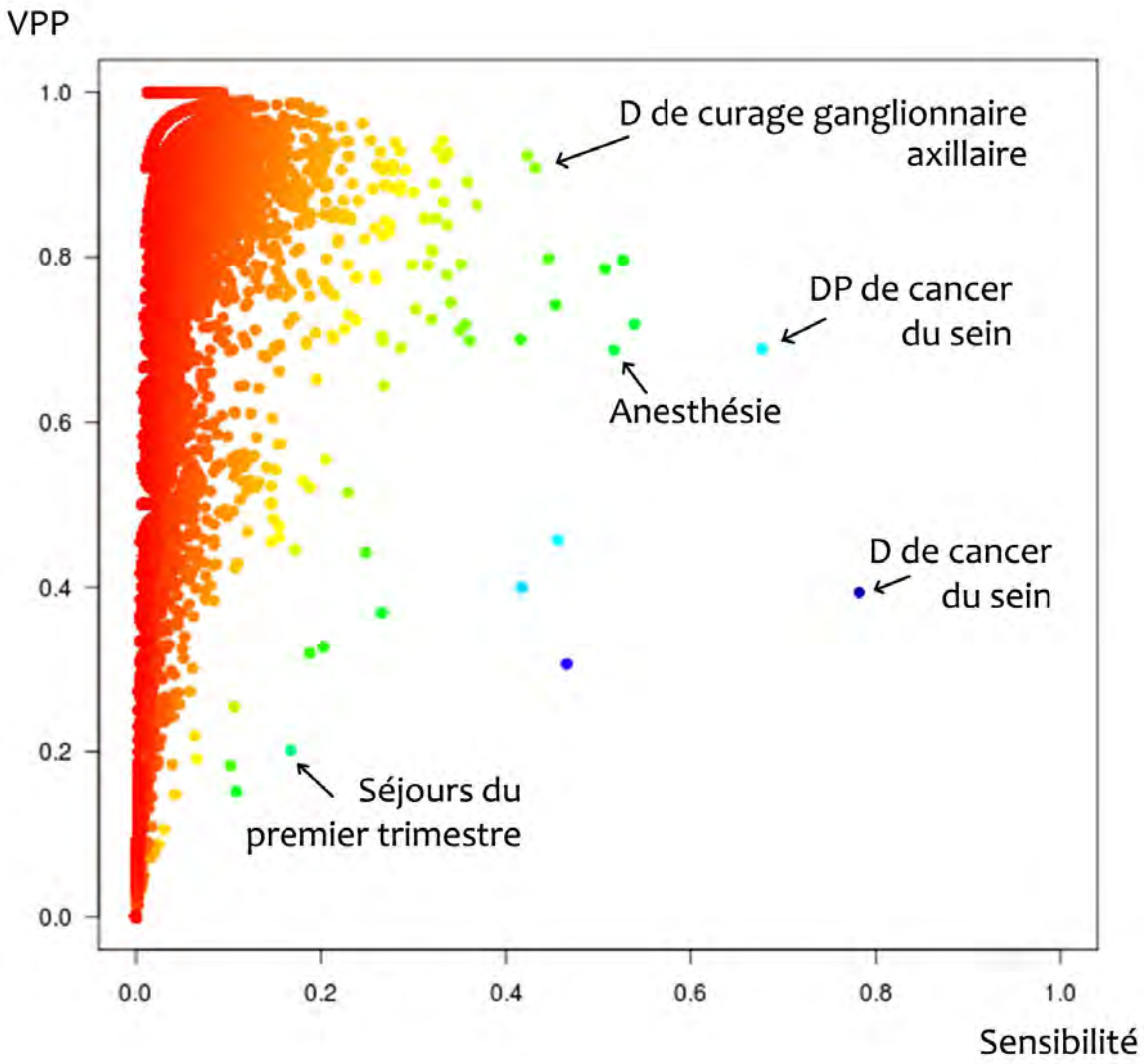

Figure 8 : Visualisation des concepts selon leur VPP et sensibilité ; la couleur représente l'effectif du concept, du plus faible (en rouge) au plus important (en bleu). Le diagnostic principal de cancer du sein semble offrir le meilleur compromis. La présence d'un curage ganglionnaire axillaire écarte de nombreux faux positifs.

Tableau 1 : Caractéristiques des concepts

Intension du concept	Effectif	Sensibilité	VPP
Concepts ayant les meilleures sensibilités			
D. de cancer du sein	1639	78,2%	39,4%
DP. de cancer du sein	810	67,6%	68,9%
D. de cancer du sein et passage en salle de réveil	618	53,8%	71,8%
DP. de cancer du sein et passage en salle de réveil	545	52,6%	79,6%
D. de cancer du sein et anesthésie générale	620	51,6%	68,7%
DP. de cancer du sein et anesthésie générale	532	50,7%	78,6%
D. de cancer du sein et âge entre 50-75	823	45,6%	45,7%
D. de cancer du sein, anesthésie générale et passage en salle de réveil	504	45,3%	74,2%
DP. de cancer du sein, anesthésie générale et passage en salle de réveil	461	44,6%	79,8%
D. de cancer du sein et curage ganglionnaire axillaire unilatéral	392	43,2%	90,8%
Concepts d'effectif supérieur à 150, ayant les meilleures VPP			
DP. de cancer du sein, curage ganglionnaire axillaire unilatéral, mastectomie partielle et passage en salle de réveil	159	18,9%	98,1%
DP. de cancer du sein, curage ganglionnaire axillaire unilatéral, mastectomie partielle, anesthésie générale et passage en salle de réveil	152	18,1%	98,0%
D. de cancer du sein, curage ganglionnaire axillaire unilatéral, mastectomie partielle, anesthésie générale et passage en salle de réveil	159	18,5%	96,2%
DP. de cancer du sein, curage ganglionnaire axillaire unilatéral et mastectomie partielle	210	24,5%	96,2%
D. de cancer du sein, curage ganglionnaire axillaire unilatéral, mastectomie partielle et passage en salle de réveil	168	19,5%	95,8%
DP. de cancer du sein, curage ganglionnaire axillaire unilatéral, mastectomie partielle et anesthésie générale	191	22,2%	95,8%
DP. de chimiothérapie, DP. de cancer du sein et mastectomie partielle	160	18,3%	94,4%
DP. de chimiothérapie, DP. de cancer du sein et curage ganglionnaire axillaire unilatéral	177	20,2%	94,4%
D. de cancer du sein, mastectomie partielle et curage	222	25,3%	94,1%

ganglionnaire axillaire unilatéral			
DP. de cancer du sein, curage ganglionnaire axillaire unilatéral et passage en salle de réveil	290	33,1%	94,1%
Concepts d'effectif supérieur à 150, ayant les VPP les plus basses			
D. de cancer du sein, D. de tumeur secondaire d'autre siège et date de sortie dans le 1er trimestre	153	1,1%	5,9%
D. de cancer du sein, D. de chimiothérapie et D. de tumeur secondaire des organes respiratoires et digestifs	191	1,6%	6,8%
DP. De chimiothérapie, D. de cancer du sein, D. de chimiothérapie et D. de tumeur secondaire des organes respiratoires et digestifs	172	1,5%	7,0%
D. de cancer du sein et D. de tumeur secondaire des organes respiratoires et digestifs	227	2,1%	7,5%
DP. de chimiothérapie, D. de cancer du sein, D. de chimiothérapie et D. de tumeur secondaire d'autres sièges	208	2,5%	10,1%
D. de cancer du sein, D. de chimiothérapie et D. de tumeur secondaire d'autres sièges	220	2,8%	10,5%
D. de cancer du sein et D. de tumeur secondaire d'autres sièges	271	4,1%	12,5%
D. de cancer du sein, D. de chimiothérapie et date de sortie dans le 1er trimestre	476	10,3%	17,9%
DP. de chimiothérapie, D. de cancer du sein, D. de chimiothérapie et date de sortie dans le 1er trimestre	458	10,2%	18,3%
D. de cancer du sein et date de sortie dans le 1er trimestre	684	16,7%	20,2%

D = Diagnostic (principal, relié ou associé) ; DP = Diagnostic principal

Discussion

Les résultats de l'AFC mettent en avant un algorithme de sélection recherchant uniquement un diagnostic principal de cancer du sein. Il s'agit de l'un des deux algorithmes les plus couramment utilisés pour identifier les cas incidents de cancer du sein dans le PMSI, le deuxième ajoutant la notion d'acte d'exérèse [16, 31, 32, 35, 42-46]. Si ce résultat n'est pas surprenant, l'analyse a permis comparer les algorithmes usuels à de nombreux autres en termes de VPP et Sensibilité.

On remarque la présence de diagnostic (et même diagnostic principal) de chimiothérapie parmi les concepts associés à une haute VPP, alors qu'on s'attend à les voir correspondre à des patientes prises en charge pour un cancer déjà connu (cas prévalents). Mais ceci peut s'expliquer par la réalisation de chimiothérapie pré- ou péri-opératoire. Ces chimiothérapies correspondent alors bien à des patients ayant un cancer du sein récemment diagnostiqué, donc des cas incidents.

Les diagnostics de tumeurs secondaires sont très présents dans les algorithmes ayant de faibles VPP. Ces diagnostics correspondent en effet le plus souvent à des cancers déjà évolués, et donc à des cas prévalents et non incidents. Un autre attribut récurrent parmi les algorithmes ayant les plus basses VPP est une date de sortie du séjour durant le premier trimestre de l'année. Ceci traduit un décalage temporel entre diagnostic du cancer et séjour PMSI à l'origine d'un effet de borne en début et fin de la période d'étude qui a déjà été décrit [16]. En effet, alors que le registre retient la date exacte du diagnostic de cancer, le PMSI n'enregistre les séjours qu'au moment de la sortie de la patiente. De plus le diagnostic de cancer peut avoir été posé avant que le patient ne soit hospitalisé pour son cancer. L'AFC a permis de détecter automatiquement ces traits caractéristiques de faux positifs.

Certains actes améliorent la VPP en réduisant le nombre de faux positifs. En revanche, la sensibilité décroit rapidement. On remarque que l'acte de mastectomie pris isolément n'est pas celui qui élimine le plus de faux positifs, alors qu'il s'agit d'un marqueur couramment utilisé pour identifier les cas incidents du cancer du sein. Les actes d'anesthésie générale et de passage en salle de réveil sont en effet plus sélectifs dans ce contexte. Ceci peut s'expliquer par l'existence de plusieurs codes de la CdAM pouvant décrire un acte d'exérèse de tumeur du sein, tandis qu'un seul code est utilisé pour l'anesthésie générale ou le passage en salle de réveil. On retrouve ici une des limites de l'AFC liée à la granularité de l'information codée dans les attributs. Un nouveau cycle

d'extraction de connaissances portant sur des regroupements d'actes permettrait de dépasser cette limite. L'association du processus d'ECD à des méthodes de représentation des connaissances est également susceptible d'améliorer la phase préparatoire des données.

L'AFC permet d'analyser les caractéristiques de nombreux algorithmes basés sur une conjonction d'attributs du PMSI. Elle résout en partie les problèmes combinatoires liés à la diversité et à la richesse des informations du PMSI. Néanmoins, la possibilité de tester des automatiquement des algorithmes reposant en partie sur des disjonctions d'attributs est une des pistes d'amélioration de notre approche.

Cette analyse présente des limites liées aux données qui ne remettent pas directement en question l'intérêt d'une approche par ECD. Premièrement, les données PMSI initiales ne concernaient que les séjours contenant un diagnostic de cancer du sein. Il n'était pas possible d'étudier les caractéristiques des faux négatifs dans le PMSI, qui étaient en grande partie déjà exclus. Une autre limite est liée à l'évolution du PMSI : depuis 2001, les pratiques médicales, les règles et nomenclatures du PMSI ont changé. Il est probable que les résultats soient partiellement reproductibles aujourd'hui. Il serait intéressant de répéter cette étude sur des données plus récentes, portant éventuellement sur une autre région disposant d'un registre. Il ne faut pas néanmoins sous-estimer la difficulté de constituer une base de données regroupant PMSI et registre du cancer, puisqu'elle nécessite une anonymisation des données du PMSI directement auprès des établissements concernés et une méthodologie de jonction complexe [52].

Au-delà de l'estimation de l'incidence du cancer du sein, les résultats de cette étude démontrent la faisabilité, la simplicité et l'intérêt de l'AFC pour l'exploitation des données du PMSI dans d'autres contextes, notamment pour des pathologies pour lesquelles les algorithmes de sélections sont moins évidents (d'autres cancers ou d'autres maladies chroniques). Cette méthode peut

également être transposée pour l'estimation d'indicateurs autres que l'incidence, pour l'inclusion de patients dans des études cliniques ou encore en matière de contrôle qualité des données.

Conclusion

L'utilisation d'une méthode de fouille de données, l'Analyse Formelle de Concepts, sur les données du PMSI a permis de construire et tester automatiquement une grande quantité d'algorithmes pour la sélection des cas incidents de cancer du sein. Cette approche montre que la recherche d'un diagnostic principal de cancer du sein présente un des meilleurs compromis entre sensibilité et VPP, au regard de très nombreux autres critères de sélection. Au-delà de ces résultats, la méthode apporte des précisions sur les limites et les biais des données du PMSI. Elle apparait prometteuse pour faciliter l'utilisation de bases médico-administratives dans le calcul d'indicateurs épidémiologiques et l'évaluation de la qualité des données.

Remerciements

Nous remercions les membres du groupe Onc-Epi (Observatoire des cancers en région Rhône Alpes) pour leur éclairage et la mise à disposition du matériel de cette étude.

CHAPITRE VI
–
Conclusion et perspectives

L'objectif de ce travail était de démontrer la faisabilité et l'intérêt de la fouille de donnée, et en particulier des treillis de Galois, pour la construction des algorithmes de sélection des cas incidents de cancer d'après les données du PMSI. Les résultats que nous avons obtenus pour le cancer du sein ont permis de confirmer que la présence d'un diagnostic principal de cancer du sein était le meilleur marqueur possible parmi l'ensemble des associations de diagnostics et d'actes pour identifier les cas incidents. Il s'agit de l'un des algorithmes déjà largement utilisés dans la littérature. Ceci s'explique car le diagnostic du cancer du sein est relativement aisé et sa prise en charge hospitalière relativement standardisée, ce qui permet de construire empiriquement des algorithmes d'identification des cas assez performants. L'abondance de la littérature sur l'utilisation du PMSI pour l'étude de ce cancer en particulier en témoigne.

Pour d'autres cancers en revanche, l'identification des cas incidents peut être plus difficile. Cooper et al [17] ont notamment montré que les données de Medicare permettaient une meilleure identification des cas de cancer du sein que des cas de cancer du colon, de l'utérus, du poumon, du pancréas et de la prostate. On peut également s'intéresser à des pathologies autres que le cancer, pour peu qu'elles impliquent une prise en charge hospitalière.

Outre la transposition à d'autres pathologies que le cancer du sein, l'utilisation des treillis de Galois devrait également se révéler plus performante pour l'exploitation de données du PMSI plus récentes. En effet, la mise en place progressive de la tarification à l'activité depuis l'année 2004 a fortement amélioré la qualité et l'exhaustivité des données du PMSI. Ces données plus complètes et

plus fiables permettront de faire apparaitre davantage de diagnostics associés ou d'actes dans les algorithmes.

Un autre avantage à l'utilisation des treillis de Galois pour la construction d'algorithmes est la possibilité d'identifier facilement un algorithme adapté à un objectif précis, selon sa sensibilité, sa spécificité et ses valeurs prédictives. L'utilisation des bases médico-administratives comme source d'information pour les registres du cancer ou pour le recrutement de patients avant inclusion dans des protocoles de recherche n'attendent en effet pas les mêmes qualités d'un algorithme d'identification des cas que l'estimation d'un taux d'incidence au niveau d'une population.

Enfin, l'utilisation des informations du chaînage anonyme pour la détection des cas incidents, qui commence à devenir possible pour le PMSI [11], va complexifier les algorithmes en permettant de tenir compte de la chronologie de prise en charge des patients. Là encore, l'utilisation des treillis de Galois permettra d'assister efficacement la construction de ces algorithmes.

BIBLIOGRAPHIE

Références

1. LOI n° 91-748 du 31 juillet 1991 portant réforme hospitalière.
2. Circulaire DHOS-F2-O / DSS-1A-2004 n° 36 du 2 février 2004 relative à la campagne budgétaire pour 2004 des établissements sanitaires financés par dotation globale.
3. Fetter RB, Shin Y, Freeman JL, Averill RF, Thompson JD. Case mix definition by diagnosis-related groups. Med Care. 1980;18:1-53.
4. Circulaire DHOS/PMSI n° 106 du 22 février 2001 relative au chaînage des séjours en établissements de santé dans le cadre du programme de médicalisation des systèmes d'information (PMSI). Bulletin officiel n°2001-13.
5. Dussaucy A, Viel JF, Mulin B, Euvrard J. L'outil PMSI : biais, sources d'erreurs et conséquences. Revue d'Epidémiologie et de Santé Publique. 1004;42:345-58.
6. Couris CM, Ecochard R. Utilisation du PMSI pour l'épidémiologie : Faisabilité, conditions d'utilisation et limites. Gestion Hospitalières. 2005;449:651-4.
7. Geoffroy-Perez B, Imbernon E, Gilg Soit Ilg A, Goldberg M. Confrontation des données du Programme nationale de surveillance du mésothéliome (PNSM) et du Proramme de médicalisation des systèmes d'information (PMSI). Rev Epidemiol Sante Publique 2006;54:475-83.
8. Milcent C, Dormont B, Durand-Zaleski I, Steg PG. Gender differences in hospital mortality and use of percutaneaous coronary intervention in acute myocardial infarction. Circulation 2007;115:833-39.

9. Hafdi-Nejjari Z, Couris CM, Schott AM, Perrot L, Bourgoin F, Borson-Chazot F. Place du Programme de médicalisation des systèmes d'information (PMSI) dans l'estimation de l'incidence du cancer de la thyroïde dans la région Rhône-Alpes. Rev Epidemiol Sante Publique 2006;54:391-8.
10. Trombert Paviot B, Gomez F, Olive F, Polazzi S, Remontet L, Bossard N, Mitton N, Colonna M, Schott AM. Identifying Prevalent Cases of Breast Cancer in the French Case-mix Databases. Methods Inf Med. 2010;49(5). [Epub ahead of print]
11. Olive F, Gomez F, Schott AM, Remontet L, Bossard N, Mitton N, Polazzi S, Colonna M, Trombert-Paviot B. Analyse critique des données du PMSI pour l'épidémiologie des cancers : une approche longitudinale devient possible. Rev Epidemiol Sante Publique. 2011. [Epub ahead of print]
12. Levi F, Bosetti C, Lucchini F, Negri E, La Vecchia C. Monitoring the decrease in breast cancer mortality in Europe. Eur J Cancer Prev 2005;14:497-502.
13. Autier P, Boniol M, La Vecchia C, Vatten L, Gavin A, Héry C, Heanue M. Disparities in breast cancer mortality trends between 30 European countries: retrospective trend analysis of WHO mortality database. BMJ. 2010;341:c3620.
14. Belot A, Grosclaude P, Bossard N, Jougla E, Benhamou E et Al. Cancer incidence and mortality in France over the period 1980-2005. Rev Epidemiol Sante Publique. 2008;56(3):159-75.
15. Comité national de pilotage. Cahier des charges du programme national de dépistage systématique du cancer du sein. DGS 1994 - mise à jour : janvier 1996.
16. Couris CM, Forêt-Dodelin C, Rabilloud M, Colin C, Bobin JY, Dargent D, Raudrant D, Schott AM. Sensibilité et spécificité de deux méthodes

d'identification des cancers du sein incidents dans les services spécialisés à partir des données médico-administratives. Rev Epidemiol Sante Publique. 2004;52(2):151-60.

17. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. Med Care. 1999;37:436-44.
18. Carre N, Uhry Z, Velten M, Tretarre B, Schvartz C, Molinie F, et al. [Predictive value and sensibility of hospital discharge system (PMSI) compared to cancer registries for thyroid cancer (1999-2000)]. Rev Epidemiol Sante Publique. 2006;54:367-76.
19. McBean AM, Babish JD, Warren JL. Determination of lung cancer incidence in the elderly using Medicare claims data American Journal of Epidemiology. 1993;137:226-34.
20. McBean AM,Warren JL, Babish JD. Measuring the incidence of cancer in elderly Americans using Medicare claims data. Cancer. 1994;73:2417-25.
21. McClish D, Penberthy L. Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. Med Care. 2004;42:1111-6.
22. McClish DK, Penberthy L, Whittemore M, Newschaffer C, Woolard D, Desch CE, et al. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. Am J Epidemiol. 1997;145(3):227-33.
23. Middleton RJ, Gavin AT, Reid JS, O'Reilly D. Accuracy of hospital discharge data for cancer registration and epidemiological research in Northern Ireland. Cancer Causes Control. 2000;11:899-905.
24. Penberthy L, McClish D, Manning C, Retchin S, Smith T. The added value of claims for cancer surveillance: results of varying case definitions. Med Care. 2005;43:705-12.
25. Penberthy L, McClish D, Pugh A, Smith W, Manning C, Retchin S. Using

hospital discharge files to enhance cancer surveillance. Am J Epidemiol. 2003;158:27-34.

26. Stang A, Glynn RJ, Gann PH, Taylor JO, Hennekens CH. Cancer occurrence in the elderly: agreement between three major data sources. Ann Epidemiol. 1999;9:60-7.
27. Whittle J, Steinberg EP, Anderson GF, Herbert R. Accuracy of Medicare claims data for estimation of cancer incidence and resection rates among elderly Americans. Med Care. 1991;29:1226-36.
28. Couris CM, Seigneurin A, Bouzbid S, Rabilloud M, Perrin P, Martin X, et al. French claims data as a source of information to describe cancer incidence: predictive values of two identification methods of incident prostate cancers. J Med Syst. 2006;30:459-63.
29. Toniolo P, Pisani P, Vigano C, Gatta G, Repetto F. Estimating incidence of cancer from a hospital discharge reporting system. Rev Epidemiol Sante Publique. 1986;34:23-30.
30. Brackley ME, Penning MJ, Lesperance ML. In the absence of cancer registry data, is it sensible to assess incidence using hospital separation records? Int J Equity Health. 2006;5:12.
31. Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. Journal of Clinical Epidemiology. 2000;53:605-14.
32. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An algorithm for the use of Medicare claims data to identify women with incident breast cancer. Health Serv Res. 2004;39:1733-49.
33. Solin LJ, MacPherson S, Schultz DJ, Hanchak NA. Evaluation of an algorithm to identify women with carcinoma of the breast. Journal of Medical Systems 1997;21:189-99.
34. Warren JL, Feuer E, Potosky AL, Riley GF, Lynch CF. Use of Medicare

hospital and physician data to assess breast cancer incidence. Med Care. 1999;37:445-56.

35. Couris CM, Schott AM, Morgon E, Ecochard R, Colin C. A literature review to assess the use of claims databases in identifying cancer incident cases. Health Services and Outcomes Research Methodology 2003;4:49-63.

36. Ganry O, Taleb A, Peng J, Raverdy N, Dubreuil A. Evaluation of an algorithm to identify incident breast cancer cases using DRGs data. Eur J Cancer Prev. 2003;12:295-9.

37. Koroukian SM, Cooper GS, Rimm AA. Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. Health Serv Res. 2003;38:947-60.

38. Leung KM, Hasan AG, Rees KS, Parker RG, Legorreta AP. Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. J Clin Epidemiol. 1999;52:57-64.

39. McGeechan K, Kricker A, Armstrong B, Stubbs J. Evaluation of linked cancer registry and hospital records of breast cancer. Aust N Z J Public Health. 1998;22:765-70.

40. Wang PS, Walker AM, Tsuang MT, Orav EJ, Levin R, Avorn J. Finding incident breast cancer cases through US claims data and a state cancer registry. Cancer Causes Control. 2001;12:257-65.

41. Warren JL, Riley GF, McBean AM, et al. Use of Medicare data to identify incident breast cancer patients. Health Care Financing Rev. 1996;18:237-46.

42. Remontet L, Mitton N, Couris CM, Iwaz J, Gomez F, Olive F, Polazzi S, Schott AM, Trombert B, Bossard N, Colonna M. Is it possible to estimate the incidence of breast cancer from medico-administrative databases? Eur J Epidemiol. 2008;23:681-8.

43. Baldi I, Vicari P, Di Cuonzo D, Zanetti R, Pagano E, Rosato R, Sacerdote C, Segnan N, Merletti F, Ciccone G. A high positive predictive value algorithm

using hospital administrative data identified incident cancer cases. J Clin Epidemiol. 2008;61:373-9.

44. Couris CM, Polazzi S, Olive F, Remontet L, Bossard N, Gomez F, Schott AM, Mitton N, Colonna M, Trombert B. Breast cancer incidence using administrative data: correction with sensitivity and specificity. J Clin Epidemiol. 2009;62:660-6.
45. Couris CM, Colin C, Rabilloud M, Schott AM, Ecochard R. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. J Clin Epidemiol. 2002;55:386-91.
46. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. Health Serv Res. 2007;42:2056-69.
47. Curtis JR, Cheng H, Delzell E, Fram D, Kilgore M, Saag K, Yun H, Dumouchel W. Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example. Med Care. 2008;46:969-75.
48. Trifirò G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salamé G, Catania MA, Salvo F, David A, Moore N, Caputi AP, Sturkenboom M, Molokhia M, Hippisley-Cox J, Acedo CD, van der Lei J, Fourrier-Reglat A; EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf. 2009 Dec;18(12):1176-84.
49. Jay N, Napoli A, Kohler F. Cancer patient flows discovery in DRG databases. Stud Health Technol Inform. 2006;124:725-30.
50. Fayyad U, Piatetsky-Shapiro G, Smyth P. The kdd process for extracting useful knowledge from volumes of data. *Communications ACM 1996*; 39:27-34.
51. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine 1996; 17(3):37-54.

52. Wille R. Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. LNCS. 2005; 3626:1-33.

53. Stewart B, Kleihues P. World Cancer Report. Oxford: Oxford University Press; 2003.

54. Baron JA, Lu-Yao G, Barrett J, et al. Internal validation of Medicare claims data. Epidemiology. 1994;5:541-4.

55. Quantin C, Fassa M, Coatrieux G, Riandey B, Trouessin G, Allaert FA. Linking anonymous databases for national and international multicenter epidemiological studies: a cryptographic algorithm. Rev Epidemiol Sante Publique. 2009;57:33-9.

56. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. Lancet Oncol. 2001;2:133-40.

VU

NANCY, le **13 avril 2011**
Le Président de Thèse

Professeur F. KOHLER

NANCY, le **13 avril 2011**
Le Doyen de la Faculté de Médecine

Professeur H. COUDANE

AUTORISE À SOUTENIR ET À IMPRIMER LA THÈSE / 3604

NANCY, le **28 avril 2011**

LE PRÉSIDENT DE L'UNIVERSITÉ DE NANCY 1
Par délégation

Madame C. CAPDEVILLE-ATKINSON

RÉSUMÉ

Introduction. L'utilisation du PMSI pour l'estimation de l'incidence des cancers se heurte à l'imperfection des critères de sélection (algorithmes) des cas, construits de façon empirique parmi les milliers de codes de diagnostics ou d'actes disponibles. Les méthodes d'extraction de connaissances à partir de données (ECD) permettent de comparer tous les algorithmes se présentant sous la forme d'une conjonction d'attributs.
Objectif. Découvrir le ou les algorithme(s) ayant les meilleures qualités (sensibilité, spécificité et valeurs prédictives) pour identifier les cas de cancer du sein incidents parmi les séjours du PMSI.
Méthode. Les courts séjours du PMSI de 2001 comprenant un diagnostic de cancer du sein ont été croisés avec les données du registre du cancer de l'Isère qui constituait le gold standard pour identifier les cas de cancer du sein incidents. Une analyse formelle des concepts a permis de calculer la sensibilité, la spécificité et les valeurs prédictives de tous les algorithmes possibles, construits avec les codes de diagnostics et d'actes, l'âge, les établissements et les dates de sortie des séjours.
Résultats. Le registre a identifié 825 patientes avec un cancer du sein incident. Le PMSI a identifié 1693 patientes ayant au moins un séjour avec cancer du sein, dont 645 étaient dans le registre. Les algorithmes les plus sensibles recherchaient un diagnostic principal de cancer du sein ou un acte d'anesthésie. Les algorithmes ayant les meilleures valeurs prédictives positives (VPP) associaient mastectomie, curage ganglionnaire, anesthésie générale, passage en salle de réveil et diagnostic principal de cancer du sein. Les diagnostics de tumeurs secondaires correspondaient aux plus basses VPP.
Conclusion. Les méthodes d'ECD ont permis de traiter automatiquement tous les algorithmes possibles. Ces résultats permettent de proposer des algorithmes ciblés en fonction de l'objectif recherché (sensibilité, spécifique ou valeur prédictive).

Data mining methods to detect incident breast cancer cases from the French medico-administrative database (PMSI): a formal concept analysis with data of year 2001 from the PMSI and from the Isère cancer registry

THÈSE : MÉDECINE SPÉCIALISÉE – ANNÉE 2011

MOTS CLEFS : Epidémiologie, cancer du sein, PMSI, fouille de données, treillis de Galois

INTITULÉ ET ADRESSE DE L'U.F.R. :

UNIVERSITÉ HENRI POINCARÉ, NANCY-1
Faculté de Médecine de Nancy
9 avenue de la Forêt de Haye
54505 VANDOEUVRE-LES-NANCY Cedex