Profils patients associés à la non conformité des décisions aux recommandations de prise en charge thérapeutique des cancers du sein : utilisation de l'analyse de concepts formels

Brigitte Séroussi, Nizar Messai, Cédric Laouénan, France Mentré, Jacques Bouaud

Profils patients associés à la non conformité des décisions aux recommandations de prise en charge thérapeutique des cancers du sein : utilisation de l'analyse de concepts formels

Brigitte Séroussi[1], Nizar Messai[2], Cédric Laouénan[3], France Mentré[3] et Jacques Bouaud[4]

[1] UPMC, UFR de Médecine, Sorbonne Universités, Paris, France; AP-HP, Hôpital Tenon, Département de Santé Publique, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIM&BIO EA3969, Bobigny, France; APREC, Paris, France
Brigitte.Seroussi@tnn.aphp.fr

[2] Université François Rabelais Tours, Laboratoire Informatique EA 6300, Tours, France
Nizar.Messai@univ-tours.fr

[3] Université Paris Diderot, Sorbonne Paris Cité, UMR_S 738, Paris, France; INSERM, UMR_S 738, Paris, France; AP-HP, Hôpital Bichat, Service de Biostatistique, Paris, France
{Cedric.Laouenan,France.Mentre}@inserm.fr

[4] AP-HP, DRCD, Paris, France ; INSERM, UMR_S 872, éq. 20, CRC, Paris, France
Jacques.Bouaud@sap.aphp.fr

Résumé : Les systèmes d'aide à la décision médicale permettent d'améliorer le suivi des recommandations de pratique clinique. OncoDoc2 est un tel système s'appuyant sur des recommandations de prise en charge du cancer du sein. Malgré son utilisation en routine lors de réunions de concertation pluridisciplinaire de sénologie, des décisions non conformes aux recommandations subsistent. L'objectif est d'utiliser l'analyse de concepts formels afin de caractériser les profils patients associés aux deux modalités de la conformité. Deux étapes de pré-traitement permettant de simplifier les données à analyser sont proposées : une réduction d'attributs par suppression de ceux non statistiquement associés à la non conformité, et un gommage sélectif de valeurs. Parmi les décisions recueillies sur 3 ans à l'hôpital Tenon, 198 concernent la reprise chirurgicale et ont été analysées. Les profils patients associés à la non conformité retrouvés sont ceux pour lesquels il n'existe pas de preuve scientifique des recommandations.

Mots-clés : Analyse de concepts formels, réduction de dimensionnalité, adhésion aux recommandations de pratique clinique, système d'aide à la décision médicale, cancer du sein

1 Introduction

Les recommandations de pratique clinique, ou *guidelines*, sont des documents de synthèse élaborés à l'attention des professionnels de santé, décrivant les prises en charge recommandées pour un ensemble de situations cliniques particulières d'une pathologie donnée. Les recommandations s'appuient sur les résultats publiés de la recherche clinique établissant ainsi leur « preuve » scientifique. Elles sont produites par les société savantes ou les agences nationales de santé afin de promouvoir les bonnes pratiques, et ainsi d'améliorer la qualité des soins. Mais, diffusées sous forme textuelle, les guidelines ont un effet limité sur les prescripteurs et leur mise en oeuvre en pratique reste insuffisante (Giguère *et al.*, 2012). Des revues systématiques de la littérature (Garg *et al.*, 2005) ont montré que les systèmes d'aide à la décision médicale (SADM) informatisés peuvent être des outils efficaces pour promouvoir le suivi des guidelines. En proposant les recommandations adaptées au cas du patient, les SADM devraient en effet théoriquement résoudre le problème de la méconnaissance des guidelines par les cliniciens. Cependant, ceci n'est pas systématiquement observé (Roshanov *et al.*, 2011; Bright *et al.*, 2012). Produire des recommandations centrées-patient lors des prises de décision ne semble « ni nécessaire, ni suffisant » pour garantir des décisions médicales conformes aux guidelines (Shiffman *et al.*, 1999).

OncoDoc2 (Séroussi *et al.*, 2001) est un SADM destiné à être utilisé lors des réunions de concertation pluridisciplinaire (RCP) de cancérologie où sont prises toutes les décisions thérapeutiques concernant la prise en charge des cancers. Pour un patient donné, caractérisé par un ensemble de critères, le système fournit les plans de soins recommandés par le référentiel de bonne pratique local (CancerEst) pour la prise en charge thérapeutique des cancers du sein non métastatiques. Après une première étude réalisée à l'hôpital Tenon (AP-HP) durant laquelle le taux de conformité des décisions de RCP au référentiel CancerEst était passé de 79 % sans intervention à 93 % après utilisation d'OncoDoc2 pour les prises de décision (Séroussi *et al.*, 2007), le système a été adopté par les membres de la RCP et a été utilisé en routine pendant près de 3 ans. Le taux de conformité est resté élevé (92 %), mais montre que sur le long terme les médecins de RCP n'ont pas systématiquement suivi les recommandations qu'ils avaient pourtant collectivement élaborées, en dépit de l'utilisation d'OncoDoc2 qui les leur rappelait au moment de la décision.

La « fouille de données », ou *data mining*, regroupe de nombreuses approches pluridisciplinaires dont le but est de découvrir automatiquement

à partir de jeux de données des régularités qui pourront être généralisées comme étant de nouvelles connaissances par des experts du domaine. Des techniques de fouille de données ont déjà été testées pour l'analyse de la conformité des décisions des médecins aux guidelines (Razavi *et al.*, 2007; Svátek *et al.*, 2004). L'analyse de concepts formels (ACF) (Ganter & Wille, 1999) est une technique particulière de data mining qui permet de dériver des relations implicites au sein d'un ensemble d'objets décrits par un ensemble d'attributs. Les données sont regroupées en unités appelées concepts formels sur la base du partage de caractéristiques entre objets. Les concepts formels sont partiellement ordonnés, formant une hiérarchie de concepts appelée treillis de concepts. La richesse des structures de treillis et leurs propriétés formelles ont motivé de nombreuses applications de l'ACF à l'analyse et à l'exploration de données dans de nombreux domaines, notamment liés à la médecine (Schnabel, 2002; Blinova *et al.*, 2003; Jay *et al.*, 2006; Endres *et al.*, 2009; Messai *et al.*, 2011).

Le travail présenté ici s'inscrit dans la recherche des causes de non conformité aux recommandations des décisions médicales. Si, de manière générale, des barrières à la mise en œuvre des guidelines ont été globalement identifiées (Cabana *et al.*, 1999), nous cherchons à caractériser plus finement les situations qui seraient associées aux décisions non conformes, en se focalisant sur les caractéristiques des patients, les autres dimensions étant ici figées : les décisions sont prises par les mêmes médecins, au sein de la même RCP, dans le même contexte hospitalier de l'hôpital Tenon, sur la base du même référentiel (CancerEst), et avec le même SADM (OncoDoc2). L'objectif est d'explorer comment l'ACF peut être utilisée pour mettre en évidence ces caractéristiques patients associées à la non conformité dans les données recueillies lors de l'utilisation d'OncoDoc2.

2 Contexte

2.1 Oncodoc2

OncoDoc2 est un SADM développé selon une approche documentaire de l'aide à la décision (Séroussi *et al.*, 2001). Son utilisation consiste à naviguer dans une base de connaissances structurée sous la forme d'un arbre décisionnel représentant le contenu des guidelines, ici le référentiel CancerEst. Les nœuds sont des variables décisionnelles catégorielles qui, pour le cancer du sein, concernent le patient, son histoire thérapeutique et les caractéristiques tumorales. Les arcs sortants correspondent aux valeurs possibles de la variable. Les recommandations de plans de soins sont

attachées aux feuilles de l'arbre. Le chemin parcouru depuis la racine représente le profil patient pour lequel ces recommandations s'appliquent. Ce profil est constitué d'un ensemble de couples <variable,valeur>.

Lors d'une utilisation pour un patient donné, les variables sont instanciées interactivement par l'utilisateur-médecin qui réalise une navigation dans la base de connaissance en sélectionnant au niveau de chaque nœud la valeur de la variable qui convient à la situation clinique à décrire, jusqu'à aboutir à une feuille où les plans de soins recommandés sont rappelés. En situation réelle, lors de l'étude d'un cas, les participants de la RCP prennent leur décision thérapeutique en pleine connaissance des recommandations que le système OncoDoc2 leur rappelle. La décision prise est dite « conforme » (aux recommandations), si elle est identique ou plus générale que au moins un plan de soins recommandé. Sinon, elle est dite « non conforme ». Pour chaque décision, on enregistre le profil patient construit lors de la navigation dans OncoDoc2 ainsi que le statut de conformité de cette décision.

2.2 Notions de base de l'ACF

Un *contexte formel* est un triplet (G, M, I), où G est un ensemble d'objets, M un ensemble d'attributs, et I une relation binaire entre les 2 ensembles indiquant si un objet de G possède ou non un attribut de M (Ganter & Wille, 1999). Un contexte formel est généralement représenté par un tableau où les lignes correspondent aux objets, les colonnes aux attributs, et où chaque case (i, j) contient ou non une croix ($\times$), selon que l'objet i est lié à l'attribut j selon la relation I.

L'ACF vise à regrouper les objets en fonction des attributs qu'ils ont en commun. Un *concept formel* est défini comme l'ensemble maximal d'objets (son extension) qui ont en commun un ensemble maximal d'attributs (son intension) par la relation I. Les concepts formels sont partiellement ordonnés sur la base de l'inclusion de leur extension, et, de manière duale, de leur intension. L'ensemble des concepts issus d'un contexte formel forme alors une hiérarchie de concepts ayant la structure d'un treillis. Les treillis de concepts peuvent être visualisés sous la forme d'un diagramme de Hasse. Pour calculer les concepts et visualiser les treillis, nous avons utilisé l'outil ConExp (http ://conexp.sourceforge.net/). Dans le mode de visualisation adopté, chaque nœud correspond à un concept dont la taille est proportionnelle à son extension. Cet affichage permet de repérer visuellement les concepts les plus fréquents. Par ailleurs, l'intension d'un concept s'obtient en cumulant tous les attributs affichés au niveau des concepts qui

le subsument. En fouille de données, les treillis de concepts sont considérés comme une représentation de différentes familles de règles d'association (Stumme *et al.*, 2002). Ainsi, chaque arête du treillis, considérée de bas en haut, correspond à une implication.

2.3 Pré-traitement classique : échelonnage plat

Dans sa forme originelle, l'ACF requiert que les données soient binaires, rassemblées dans un contexte formel, ce qui est rarement le cas des données réelles. Une étape de pré-traitement, ou échelonnage conceptuel, est alors nécessaire pour transformer les données en un contexte formel.

Dans notre application, les objets correspondent aux décisions de RCP, et les attributs correspondent aux variables. Un ensemble de décisions, telles que celles enregistrées lors de l'utilisation d'OncoDoc2, peut être représenté sous forme tabulaire (tableau 1) où les lignes représentent les décisions D_i, les colonnes représentent les variable V_j (critères cliniques et statut de conformité) et où une case contient pour une décision donnée la valeur de la variable instantiée au cours de la navigation. Les variables sont catégorielles, souvent binaires. De par la structure arborescente de la base de connaissances d'OncoDoc2, les décisions ne partagent pas les mêmes variables et les valeurs de ces variables apparaissent comme « manquantes » dans le jeu de données. Il s'agit des variables non nécessaires à la décision dans certaines situations. Elles sont représentées dans le tableau par des cases vides grisées et rendent le jeux de données incomplet.

TABLE 1 – Exemple d'un jeu de données fictif

	V_1	V_2	V_3	V_4	V_5	Compliance
D_1	A	C		G		Yes
D_2	B			G	H	Yes
D_3	A	C		G	H	Yes
D_4	B	D		G		No
D_5		E		G		No
D_6	A	C		G		Yes
D_7	B			G		Yes
D_8		E		G	H	No
D_9	B	D		G	H	No
D_{10}		C, E		G		No

Le jeu de données obtenu par l'utilisation d'OncoDoc2 est transformé en un contexte formel par l'application d'un échelonnage classique dit plat.

Celui-ci consiste à transformer chaque variable en autant d'attributs binaires que de valeurs possibles de la variable. Dans ce processus d'échelonnage, les valeurs manquantes ne sont pas considérées comme une modalité particulière pertinente et ne produisent donc pas d'attribut supplémentaire. Par exemple, sur la base du jeu de données du tableau 1, la variable binaire V_1 prend les 2 valeurs A et B. Elle est transformée en 2 attributs binaires '$V_1 = A$' et '$V_1 = B$', notés respectivement A et B. Cet échelonnage plat appliqué au jeu de données du tableau 1 produit le contexte formel figurant dans le tableau 2. Il existe d'autres sortes d'échelonnages (Ganter & Wille, 1999; Endres *et al.*, 2009), dépendants de l'interprétation des données et des besoins applicatifs.

TABLE 2 – Contexte formel obtenu par échelonnage plat

	V_1=A (A)	V_1=B (B)	V_2=C (C)	V_2=D (D)	V_2=E (E)	V_3 (F)	V_4=G (G)	V_5=H (H)	Compl.=Yes (COMP;YES)	Compl.=No (COMP;NO)
D_1	×		×				×		×	
D_2		×					×	×	×	
D_3	×		×				×	×	×	
D_4		×		×			×			×
D_5					×		×			×
D_6	×		×				×		×	
D_7		×					×		×	
D_8					×		×	×		×
D_9		×		×			×	×		×
D_{10}			×		×		×			×

3 Méthode

3.1 Rationnel de la transformation des données

Classiquement, l'ACF est réalisée sur le contexte formel obtenu après échelonnage plat. Toutefois, le treillis obtenu peut être « touffu », ne permettant pas de distinguer les concepts formels associés à la conformité de ceux associés à la non conformité. En effet, un nombre élevé de concepts et d'arrêtes résulte de la dispersion dans les décisions des valeurs des attributs au regard de la conformité. La figure 1 montre ainsi le treillis produit par ACF sur le contexte formel du tableau 2.

Des transformations des données permettent de simplifier les contextes formels et d'obtenir des résultats d'ACF moins complexes. Ainsi, certains

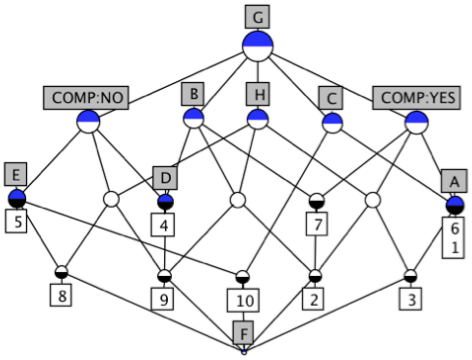

FIGURE 1 – Treillis de concepts issu du contexte formel du tableau 2

pré-traitements peuvent permettre de mieux identifier les relations entre concepts formels et modalités de conformité. Dans un travail préalable, Messai *et al.* (2011) avaient simplifié le contexte formel en éliminant les attributs représentés à la fois dans les décisions conformes et non conformes, ne gardant que les attributs associés systématiquement à l'une ou l'autre modalité de conformité afin d'obtenir 2 sous-treillis de concepts. Ainsi, même lorsque presque toutes les valeurs d'un attribut étaient associées à une modalité de la conformité, une seule valeur de cet attribut présent dans une décision de conformité opposée conduisait à la suppression de l'attribut. Cette méthode présente l'inconvénient d'éliminer beaucoup d'attributs, et d'appauvrir ainsi les données de façon excessive. Dans cet esprit, nous proposons ici une méthode de transformation des données, en prétraitement de l'ACF, ayant pour objectif de repérer les concepts associés à la non conformité ou à la conformité. Elle repose sur 2 étapes. La première étape relève d'une réduction de dimensionnalité, souvent utilisée en analyse de données, et consiste en une sélection d'attributs qui se fonde sur des critères statistiques. La seconde étape consiste à transformer les données par un gommage sélectif de valeurs.

3.2 Réduction statistique d'attributs

Une étape préliminaire, effectuée sur le contexte formel obtenu après échelonnage plat, a consisté à éliminer les attributs considérés comme non contributifs sur des critères de fréquence. Ainsi, les attributs non contributifs sont (i) ceux à valeur constante sur tout le jeu de données, p. ex. G dans l'exemple du tableau 2, ou (ii) ceux contenant « trop » de valeurs

manquantes, le seuil étant fixé arbitrairement à 95 % de la taille du jeu de données, p. ex. F dans l'exemple.

La première étape consiste à éliminer les attributs dont la distribution dans les 2 groupes de décisions, conformes et non conformes, est « équilibrée ». Pour déterminer si la distribution observée dans le jeu de données est équilibrée, nous avons employé le test de Fisher exact. En effet, ce test ne repose sur aucune hypothèse concernant la loi de distribution des variables et reste exact même pour de faibles effectifs, ce qui peut être le cas du fait des valeurs manquantes. Ainsi, on teste l'association entre chaque attribut et la conformité à partir d'un tableau de contingence. La *p-value* est calculée. Elle représente la probabilité d'observer cette distribution, ou des distributions plus extrêmes, sous l'hypothèse d'indépendance entre les 2 attributs. À partir du contexte formel, on construit pour chaque attribut le tableau de contingence sans tenir compte des cases vides des attributs correspondant à des valeurs manquantes avant échelonnage, puis on calcule la p-value. Le tableau 3 montre le contexte formel (après suppression préalable des attributs G et F) où les cases correspondant aux valeurs manquantes sont grisées et où les p-values figurent dans la dernière ligne. Par exemple, le tableau de contingence entre l'attribut C et la conformité figure ci-dessous. La p-value vaut 0,143.

	C:×	C:∅
COMP:YES	3	0
COMP:NO	1	4

TABLE 3 – Contexte formel du tableau 2 avec identification des valeurs manquantes et p-values calculées pour le test de Fisher exact

	A	B	C	D	E	H	COMP:YES	COMP:NO
D_1	×		×				×	
D_2		×				×	×	
D_3	×		×			×	×	
D_4		×		×				×
D_5					×			×
D_6	×		×				×	
D_7		×					×	
D_8					×	×		×
D_9		×		×		×		×
D_{10}			×		×			×
p-value	*0.429*	*0.429*	***0.143***	*0.464*	***0.196***	*1*	-	-

Habituellement, lors d'un test statistique, la p-value est comparée à un seuil de significativité en deça duquel on considère que l'hypothèse d'indépendance peut être rejetée et que l'association entre les 2 attributs est probable. Par la suite, nous avons utilisé 2 seuils de significativité : 5 %, valeur communément utilisée en statistique (risque de première espèce), et 20 %, moins restrictif, utilisé pour sélectionner les variables candidates pour une analyse multivariée. À l'issue de cette étape, tous les attributs qui ne sont pas significativement associés, pour un seuil donné, à la conformité, sont considérés comme non contributifs pour expliquer le statut de conformité et sont supprimés. Le contexte formel est ainsi réduit par abandon d'attributs, le nombre d'objets/décisions restant inchangé. Dans l'exemple du tableau 3 et avec un seuil à 20 %, seuls les attributs C et E seront conservés. La figure 2 montre le contexte réduit et le treillis correspondant.

	C	E	COMP:YES	COMP:NO
D_1	×		×	
D_2			×	
D_3	×		×	
D_4				×
D_5		×		×
D_6	×		×	
D_7			×	
D_8		×		×
D_9				×
D_{10}	×	×		×

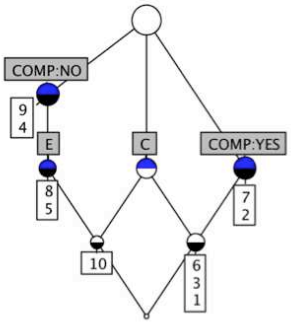

FIGURE 2 – Contexte formel du tableau 3 après réduction statistique d'attributs au seuil 20 % et treillis correspondant

3.3 Gommage sélectif de valeurs

Le contexte formel réduit lors de l'étape précédente est alors constitué d'attributs « déséquilibrés », c.à-d. distribués de façon statistiquement non uniforme entre décisions conformes et non conformes. Certains sont associés à la conformité, p. ex. C, d'autres à la non confomité p. ex. E. Afin de renforcer ces associations pour l'ACF, on considère que les valeurs d'un attribut présentes dans les décisions de conformité différente de celle à laquelle l'attribut est statistiquement associé sont du « bruit ». Ces valeurs sont alors gommées du contexte formel. On obtient alors 2 sous-treillis de concepts, l'un associé à la conformité et l'autre à la non conformité. Dans

l'exemple, avec l'attribut C qui est significativement associé à la conformité, la valeur de C pour la décision non conforme D_{10} est effacée sans toucher à celles des décisions conformes D_1, D_3 et D_6. Cette étape de gommage sélectif de valeurs produit un nouveau contexte formel simplifié où chaque attribut a des valeurs dans un seul sous-ensemble de décisions. La figure 3 montre le contexte obtenu après cette étape et le treillis correspondant ; la valeur supprimée est matérialisée par une barre oblique.

	C	E	COMP:YES	COMP:NO
D_1	×		×	
D_2			×	
D_3	×		×	
D_4				×
D_5		×		×
D_6	×		×	
D_7			×	
D_8		×		×
D_9				×
D_{10}		×		×

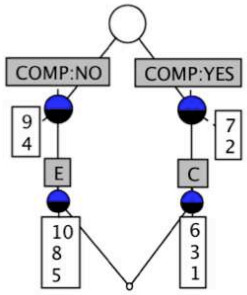

FIGURE 3 – Contexte formel de la figure 2 après gommage sélectif de valeurs et treillis correspondant

4 Résultats

Le jeux de données utilisé est constitué de 1 889 décisions de prise en charge d'un cancer du sein enregistrées entre février 2007 et septembre 2009. Trois groupes ont été constitués : les décisions pré-chirurgicales, les décisions avec indication de reprise chirurgicale et les décisions adjuvantes (c.à-d. post-chirurgicales). Pour cette étude, seul le groupe « reprise » a été traité. Ce groupe est composé de 198 décisions, parmi lesquelles 28 (14,1 %) sont non conformes au référentiel CancerEst, conduisant à un taux de conformité de 85,9 % pour ce groupe. 52 variables ont été utilisées, principalement binaires, p. ex. multifocalité = oui/non, mais parfois n-aires, p. ex. grade SBR = 1/2/3.

4.1 Échelonnage plat

Après échelonnage plat du jeu de données initial, 77 attributs ont été créés : 75 attributs patients et 2 attributs pour les 2 modalités de la confor-

mité. L'ACF sur ce contexte produit un total de 46 949 concepts. Le treillis correspondant n'a pu être tracé par ConExp.

4.2 Réduction statistique d'attributs

L'étape préliminaire a supprimé 12 attributs, 1 attribut constant et 11 attributs quasi-vides. Avec un seuil de 95 %, les attributs quasi-vides correspondent à ceux ayant moins de 10 valeurs dans le jeu des 198 décisions. Ces attributs sont issus de variables décisionnelles utilisées dans des situations très particulières et rares, p. ex. découverte d'un site invasif après chirurgie pour cancer in-situ. On notera que les décisions pour ces quelques cas étaient plutôt conformes.

Le test de Fisher exact a été réalisé pour les 65 attributs restants et pour les 2 seuils de significativité (5 % et 20 %), produisant 2 contextes réduits différents. Au seuil de 20 %, 18 attributs considérés comme significativement associés à la conformité ont été conservés. L'ACF produit 473 concepts et un treillis qui reste difficilement exploitable. Au seuil de 5 %, seuls 7 attributs sont conservés. On notera que ces attributs sont nécessairement inclus dans les 18 attributs conservés au seuil 20 %. L'ACF génère un treillis de 33 concepts, plus lisible. La figure 4 montre les 2 treillis obtenus.

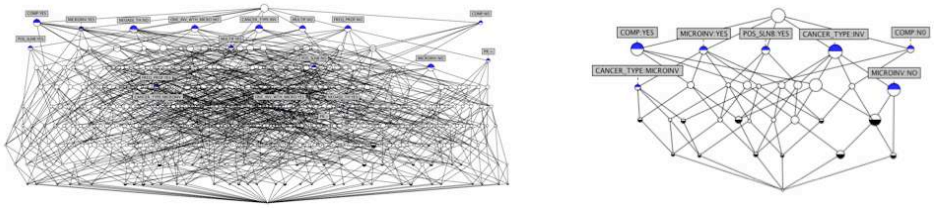

FIGURE 4 – Treillis obtenus sur le jeu de 198 décisions après réduction statistique d'attributs pour les seuils de significativité de 20 % et 5 %

4.3 Gommage sélectif de valeurs

Cette dernière étape est réalisée sur le contexte réduit lors de l'étape précédente et pour les 2 seuils de significativité. Au seuil 20 %, sur les 16 attributs cliniques dont la distribution a été modifiée par gommage de valeurs, 8 sont associés à la non conformité, et 8 sont associés à la conformité. Après gommage, l'ACF est réalisée et produit 71 concepts. Le

treillis obtenu contient 2 sous-treillis dont les racines sont {COMP:YES} et {COMP:NO}. Il y a 20 concepts associés à la non conformité et 46 concepts associés à la conformité. Au seuil de 5 %, 5 attributs ont été modifiés, 2 en faveur de la non conformité et 3 en faveur de la conformité. L'ACF produit 11 concepts, dont 2 sont associés à la non conformité et 5 à la conformité. La figure 5 donne les treillis obtenus pour les 2 seuils.

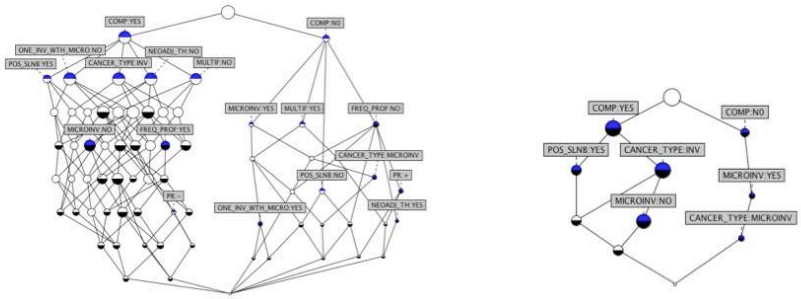

FIGURE 5 – Treillis obtenus sur le jeu de 198 décisions après réduction et gommage pour les seuils de significativité de 20 % et 5 %

5 Discussion

En dépit d'un nombre relativement faible de données (< 200 décisions), mais avec 77 attributs après échelonnage plat, les résultats de l'ACF classique ne sont pas interprétables. Le nombre élevé de concepts produits (46 949) suggère une grande dispersion des valeurs des variables dans les décisions. Afin de réduire cette complexité, nous avons proposé un pré-traitement en deux étapes des données. La première étape consiste à réaliser une réduction statistique d'attributs afin de supprimer les attributs non significativement associés à la conformité. La seconde étape de pré-traitement transforme les données en gommant de façon sélective les valeurs, considérées comme du bruit, qui ne permettent pas de traduire qualitativement, via l'ACF, les associations statistiques trouvées avec les modalités de la conformité. On obtient alors 2 sous-treillis spécifiques de chaque modalité de la conformité.

Avec le seuil de significativité de 20 %, les profils patients associés à la non conformité concernent la prise en charge des profils cliniques rares, avec des tumeurs microinvasives multifocales, dont le cas de la tumeur invasive unique associée à un foyer microinvasif, et les situations

avec ganglion sentinelle négatif, récepteurs à la progestérone positifs ou un traitement néoadjuvant antérieur. Dans ces cas, la reprise chirurgicale est recommandée mais n'est pas décidée. Pour ce qui est des tumeurs microinvasives, ceci pourrait s'expliquer par la divergence entre le référentiel local (CancerEst) et les recommandations nationales. En effet, le premier considère que les tumeurs microinvasives doivent être traitées comme des tumeurs invasives, et les secondes considèrent qu'elles doivent être traitées comme des tumeurs in-situ, bien qu'il n'y ait aucune preuve scientifique pour étayer l'une ou l'autre option. D'autres décisions non conformes concernent le cas où les cliniciens ne décident pas d'un curage axillaire après ganglion sentinelle car ils considèrent que la reprise chirurgicale n'est pas le meilleur choix pour la patiente. Au seuil de significativité de 5 %, on retrouve les critères les plus fortement associés à la non-conformité (microinvasion). Sur ce jeu de données, le seuil de 5 % est trop restrictif pour que l'ACF fournisse des résultats intéressants. Le seuil de 20 % en revanche semble un bon compromis.

Du fait de la structure arborescente de la base de connaissances d'OncoDoc2, les profils patients construits par les navigations sont constitués de jeux de variables différents. Ainsi, les objets créés contiennent des attributs avec valeurs « manquantes ». Les données réelles sont souvent incomplètes et des solutions ont été proposées pour tenir compte des données manquantes (Krupka & Lastovicka, 2012; Liu & Yao, 2010). Par ailleurs, l'échelonnage plat constitue une façon de gérer les données manquantes dans le jeu de données initial. Ainsi, les concepts produits par ACF sont identiques, que les cases vides des attributs proviennent d'une donnée manquante (case grisée) ou d'une donnée négative. Néanmoins, cette distinction a été prise en compte afin de garantir un calcul correct des p-values pour les tests d'asssociation de chaque attribut avec la conformité.

La méthode de réduction statistique d'attributs proposée ici appartient aux techniques plus générales de réduction de dimension utilisées en fouille de données pour les grands jeux de données. D'autres approches de réduction automatique d'attributs se basent sur les ensembles approximatifs (Liu *et al.*, 2007), les ensembles flous (Kang *et al.*, 2012), ou encore la quantité d'information (Xu *et al.*, 2012). La stratégie développée dans ce travail apparaît comme une manière ad-hoc de contourner les problèmes généraux liés à la prise en compte de larges jeux de données par l'ACF, qui constituent par ailleurs une question ouverte. En effet, les treillis conceptuels issus de grand jeux de données sont difficilement exploitables dès que le nombre de concepts dépasse quelques centaines (Codocedo *et al.*, 2011).

Le post-traitement des treillis et/ou des concepts une fois créés est une alternative aux transformations des données et à la perte d'information des stratégies de réduction de dimension et constitue un autre type de solution pour contrôler la complexité des treillis produits par l'ACF. Plusieurs mesures pour filtrer les concepts « intéressants » ont été définies et sont basées sur leur fréquence (Stumme *et al.*, 2002), leur stabilité (Kuznetsov, 2007), etc. afin de produire des treillis interprétables. Ce type d'approche devrait être testé sur nos données.

6 Conclusion

Ce travail exploratoire vise à utiliser l'ACF pour identifier les profils patients associés à la non conformité des décisions de RCP en dépit du rappel des recommandations par un SADM dans la prise en charge du cancer du sein. Nous avons proposé un pré-traitement en 2 étapes à une ACF classique. Il a été appliqué sur un jeu de données correspondant à des décisions thérapeutiques en fonction de la conformité ou non à des guidelines. Les profils patients liés à la non conformité sont des profils rares, faisant intervenir la présence d'une tumeur invasive avec de la microinvasion, ou un ganglion sentinelle négatif. Ces profils patients correspondent à des situations clinique pour lesquelles il n'y a pas de preuve scientifique pour garantir une prise en charge recommandée, ce qui autorise les cliniciens de RCP à être en désaccord avec leur propre référentiel. Ceci peut expliquer pourquoi, ils décident de pas toujours appliquer les recommandations malgré leur rappel par le SADM OncoDoc2. Ces profils patients représentent des profils à risque de non conformité, qui devraient être signalés aux médecins de façon à être traités en début de RCP. Ces profils sont également des candidats à l'élaboration d'essais cliniques afin de déterminer les meilleures prises en charge.

Afin de mieux comprendre l'effet patient sur la non conformité, il est certainement nécessaire de collecter de plus grands jeux de données que celui que nous avons utilisé, incluant un nombre significatif de décisions non conformes. De plus, dans le contexte de l'ACF, des méthodes alternatives à la réduction d'attributs basées sur le filtrage des concepts en post-traitement devraient permettre d'éviter la perte d'information. L'emploi en post-traitement du test d'association statistique entre l'appartenance à un concept et la conformité nous semble une piste à explorer. Par ailleurs, d'autres méthodes de fouille de données que l'ACF pourraient être testées.

Références

BLINOVA V. G., DOBRYNIN D. A., FINN V. K., KUZNETSOV S. O. & PANKRATOVA E. S. (2003). Toxicology analysis by means of the JSM-method. *Bioinformatics*, **19**(10), 1201–1207.

BRIGHT T. J., WONG A., DHURJATI R., BRISTOW E., BASTIAN L., COEYTAUX R. R., SAMSA G., HASSELBLAD V., WILLIAMS J. W., MUSTY M. D., WING L., KENDRICK A. S., SANDERS G. D. & LOBACH D. (2012). Effect of clinical decision-support systems : A systematic review. *Ann Intern Med*, **157**(1), 29–43.

CABANA M. D., RAND C. S., POWE N. R., WU A. W., WILSON M. H., ABBOUD P.-A. C. & RUBIN H. R. (1999). Why don't physicians follow clinical practice guidelines ? a framework for improvement. *JAMA*, **282**(15), 1458–1465.

CODOCEDO V., TARAMASCO C. & ASTUDILLO H. (2011). Cheating to achieve formal concept analysis over a large formal context. In *Proc. CLA11*, p. 349–362.

ENDRES D., FÖLDIÁK P. & PRISS U. (2009). An application of formal concept analysis to semantic neural decoding. *Ann Math and Artif Intell*, **57**(3-4), 233–248.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis*. Mathematical Foundations. Springer.

GARG A. X., ADHIKARI N. K. J., MCDONALD H., ROSAS-ARELLANO M. P., DEVEREAUX P. J., BEYENNE J., SAM J. & HAYNES R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes : a systematic review. *JAMA*, **293**(10), 1223–1238.

GIGUÈRE A., LÉGARÉ F., GRIMSHAW J., TURCOTTE S., FIANDER M., GRUDNIEWICZ A., MAKOSSO-KALLYTH S., WOLF F., FARMER A. & GAGNON M. (2012). Printed educational materials : effects on professional practice and health care outcomes. *Cochrane Database of Systematic Reviews*. 10 :CD004398.

JAY N., NAPOLI A. & KOHLER F. (2006). Cancer patient flows discovery in drg databases. In A. HASMAN, R. HAUX, J. VAN DER LEI & E. D. F. ROGER-FRANCE, Eds., *Ubiquity : Technologies for Better Health in Aging Societies - Proceedings of MIE2006*, volume 124 of *Stud Health Technol Inform*, p. 725–730 : IOS Press.

KANG X., LI D., WANG S. & QU K. (2012). Formal concept analysis based on fuzzy granularity base for different granulations. *Fuzzy Sets and Systems*, **203**, 33–48.

KRUPKA M. & LASTOVICKA J. (2012). Concept lattices of incomplete data. In SPRINGER, Ed., *Proc. ICFCA*, number 7278 in LNAI, p. 180–194.

KUZNETSOV S. (2007). On stability of a formal concept. *Ann Math and Artif*

Intell, **49**, 101–115.

LIU J. & YAO X. (2010). Formal concept analysis of incomplete information system. In *Proc. FSKD*, number 5, p. 2016–2020.

LIU M., SHAO M., ZHANG W. & WU C. (2007). Reduction method for concept lattices based on rough set theory and its application. *Comput Math Appl*, **53**(9), 1390–1410.

MESSAI N., BOUAUD J., AUFAURE M.-A., ZELEK L. & SÉROUSSI B. (2011). Using formal concept analysis to discover patterns of non-compliance with clinical practice guidelines : a case study in the management of breast cancer. In M. PELEG, C. COMBI, A. ABU-HANNA & S. ANDREASSEN, Eds., *AIME 2011*, Lecture Notes in Computer Science, p. 119–128 : Springer.

RAZAVI A. R., GILL H., ÅHLFELDT H. & SHAHSAVAR N. (2007). A data mining approach to analyze non-compliance with a guideline for the treatment of breast cancer. In K. A. KUHN, J. R. WARREN & T.-Y. LEONG, Eds., *MedInfo*, volume 129 of *Studies in Health Technology and Informatics*, p. 591–595 : IOS Press.

ROSHANOV P., MISRA S., GERSTEIN H., GARG A., SEBALDT R., MACKAY J., WEISE-KELLY L., NAVARRO T., WILCZYNSKI N., HAYNES R. & CCDSS SYSTEMATIC REVIEW TEAM (2011). Computerized clinical decision support systems for chronic disease management : A decision-maker-researcher partnership systematic review. *Implement Sci*, **6**(92).

SCHNABEL M. (2002). Representing and processing medical knowledge using formal concept analysis. *Methods Inf Med*, **41**(2), 160–167.

SHIFFMAN R. N., LIAW Y., BRANDT C. A. & CORB G. J. (1999). Computer-based guideline implementation systems : a systematic review of functionality and effectiveness. *JAMIA*, **6**(2), 104–114.

STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N. & LAKHAL L. (2002). Computing iceberg concept lattices with titanic. *Data & Knowledge Engineering*, **42**, 189–222.

SVÁTEK V., RÍHA A., PELESKA J. & RAUCH J. (2004). Analysis of guideline compliance–a data mining approach. *Stud Health Technol Inform*, **101**, 157–61.

SÉROUSSI B., BOUAUD J. & ANTOINE É.-C. (2001). OncoDoc, a successful experiment of computer-supported guideline development and implementation in the treatment of breast cancer. *Artif Intell Med*, **22**(1), 43–64.

SÉROUSSI B., BOUAUD J., GLIGOROV J. & UZAN S. (2007). Supporting multidisciplinary staff meetings for guideline-based breast cancer management : a study with OncoDoc2. In *Proc AMIA 2007*, p. 656–660, Chicago, IL : AMIA.

XU E., YANG Y. & REN Y. (2012). A new method of attribute reduction based on information quantity in an incomplete system. *Journal of Software*, **7**(8), 1881–1888.