Identification des profils de changement sur données longitudinales, illustrée par deux exemples : étude des trajectoires hopsitalières de prise en charge d'un cancer. Construction des profils évolutifs de qualité de vie lors d'un essai thérapeutique pour un cancer avancé

Gilles Eric Nuemi Tchathouang

UNIVERSITE DE BOURGOGNE

École Doctorale Environnement-Santé
Centre d'Épidémiologie des Populations

IDENTIFICATION DES PROFILS DE CHANGEMENT SUR DES DONNÉES LONGITUDINALES, ILLUSTRÉE PAR DEUX EXEMPLES :

- ***ÉTUDE DES TRAJECTOIRES HOSPITALIÈRES DE PRISE EN CHARGE D'UN CANCER.***
- ***CONSTRUCTION DES PROFILS ÉVOLUTIFS DE QUALITÉ DE VIE LORS D'UN ESSAI THÉRAPEUTIQUE POUR UN CANCER AVANCÉ.***

Thèse présentée pour
l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE BOURGOGNE

Discipline : Santé Publique / Épidémiologie

par Gilles Éric NUEMI TCHATHOUANG

Le 21 octobre 2014

Composition du Jury

Professeur François KOHLER	Rapporteur
Professeur Jean-Luc NOVELLA	Rapporteur
Professeur Christine BINQUET	Examinateur
Professeur Catherine QUANTIN	Directeur

REMERCIEMENTS

À l'Institut national du Cancer (INCa) pour le financement du projet discuté dans la première partie :

Étude des trajectoires de prise en charge des cancers dans la région bourgogne :

Application aux cancers du sein, du côlon-rectum et du poumon.

Au conseil régional de Bourgogne pour le financement du projet discuté dans la deuxième partie :

Effet du traitement et des caractéristiques des patients atteints d'un cancer digestif sur le profil évolutif de leur qualité de vie.

Et, qui a permis le travail sur la construction des profils évolutifs de qualité de vie

À la fédération française de cancérologie digestive (FFCD) et ses membres de l'Unité INSERM, U866 de l'université de Bourgogne,

Pour la mise à disposition des données de l'essai clinque sans lesquelles le travail sur les profils évolutifs de qualité de vie n'aurait pas été possible

À l'association du prix Liliane et Pierre Dusserre,

Qui a soutenu ce travail par le Prix Liliane Dusserre 2012 sous la forme de bourse.

À mon directeur de thèse,

Pour son dévouement, ses encouragements et son soutien durant tout ce travail.

À toute ma famille,

Pour leur patience tout au long de ce travail.

Au Professeur François KOHLER,

Qui nous a fait l'honneur de partager son avis sur ce travail et de participer au jury de thèse.

Puissiez-vous trouver ici l'expression de notre profond respect

Au Professeur Jean-Luc NOVELLA,

Qui a accepté d'être rapporteur de ce travail. C'est un honneur de vous compter parmi les membres du jury.

Puissiez-vous trouver ici le témoignage de notre profond respect

Au Professeur Christine BINQUET,

Pour votre disponible, votre patience et votre accompagnement.

Au Professeur Francis GUILLEMIN,

Pour votre disponible, votre patience et votre accompagnement.

Résumé

Contexte

Dans le domaine de la santé, l'analyse des données pour l'extraction des connaissances est un enjeu en pleine expansion. Les questions sur l'organisation des soins ou encore l'étude de l'association entre le traitement et qualité de vie (QdV) perçue pourraient être abordées sous cet angle. L'évolution des technologies permet de disposer d'outils de fouille de données performants et d'outils statistiques enrichis de méthode avancées, utilisables par les non-experts. Nous avons illustré cette méthode au travers de deux questions d'actualité :1 / Quelle organisation des soins pour la prise en charge des cancers ? 2/ étude de la relation chez les patients souffrant d'un cancer métastatique entre la QdV liée à la santé perçue et les traitements reçus dans le cadre d'un essai thérapeutique.

Matériels et méthodes

Nous disposons aujourd'hui de volumineuses bases de données. Certaines retracent le parcours hospitalier des patients, comme c'est le cas pour les données d'activités hospitalières recueillies dans le cadre du programme de médicalisation des systèmes d'information (PMSI). D'autres conservent les informations sur la QdV perçues par les patients et qui recueillies en routine actuellement dans les essais thérapeutiques. L'analyse de ces données a été réalisée suivant trois étapes principales : Tout d'abord une étape de préparation des données dont l'objectif était la compatibilité à un concept d'analyse précisé. Il s'agissait par exemple de transformer une base de données classique (centrée sur le patient) vers une nouvelle base de données où « l'unité de recueil » est une entité autre que le patient (ex. trajectoire de soins). Ensuite une deuxième étape consacrée à l'application de méthodes de fouille de données pour l'extraction connaissances : les méthodes d'analyse formelle des concepts ou encore les méthodes de classifications non-supervisée. Et enfin l'étape de restitution des résultats obtenus et présenté sous forme graphique.

Résultats

Pour la question de l'organisation des soins, nous avons construit une typologie des trajectoires hospitalières des soins permettait de réaliser un état des lieux des pratiques dans la prise en charge des cancers étudié depuis la chirurgie jusqu'à un an de suivi des patients. Dans le cas du Cancer du sein, nous avons décrit une typologie de prise en charge sur la base des coûts d'hospitalisation sur un suivi d'un an. Pour la deuxième question, nous avons également construit une typologie des profils évolutifs de la QdV. Celle-ci comportait 3 classes : une classe d'amélioration, une classe de stabilité et une classe de dégradation.

Conclusion

L'intérêt majeur de ce travail était de mettre en évidence des pistes de réflexion permettant des avancées dans la compréhension et la construction de solutions adaptées aux problèmes.

Mots clés

Fouille de données ; Classification ; Cancers ; trajectoire de soins ; Qualité de vies ; Imputation de données.

Abstract

Context

In healthcare domain, data mining for knowledge discovery represent a growing issue. Questions about the organisation of healthcare system and the study of the relation between treatment and quality of life (QoL) perceived could be addressed that way. The evolution of technologies provides us with efficient data mining tools and statistical packages containing advanced methods available for non-experts. We illustrate this approach through two issues: 1 / What organisation of healthcare system for cancer diseases management? 2 / Exploring in patients suffering from metastatic cancer, the relationship between health-related QoL perceived and treatment received as part of a clinical trial.

Materials and methods

Today we have large databases. Some are dedicated to gather together all hospital stays, as is the case for the national medico-administrative DRG-type database. Others are used to store information about QoL perceived by patients, routinely collected in clinical trials. The analysis of these data was carried out following three main steps: In the first step, data are prepared to be useable according to a defined concept of data analysis. For example, a classical database (patient-centered) was converted to a new database organised around a new defined entity which was different from the patient (eg. Care trajectory). Then in the second step, we applied data mining methods for knowledge discovery: we used the formal analysis of concepts method and unsupervised clustering techniques. And finally the results were presented in a graphical form.

Results

Concerning the question of the organisation of healthcare system, we constructed a typology of hospital care trajectories. We were able then to describe current practice in the management of cancers from the first cancer related surgical operation until one year of follow-up. In the case of breast cancer, we've described a typology of care on the basis of hospital costs over a one year follow up. Concerning the second question, we have also constructed a typology of QoL change patterns. This comprised three groups: Improvement, stability and degradation group.

Conclusion

The main interest of this work was to highlight new thoughts, which advances understanding and, contributing in appropriate solutions building.

Keywords :

Data mining; Clustering; Cancer; Trajectory of care; Quality of life; Multiple imputation

TABLE DES MATIÈRES

INTRODUCTION

Dans le domaine de la santé, le contexte économique et social induit des problématiques nouvelles. Ainsi, la question du mode d'organisation des soins à mettre en place pour une meilleure prise en charge des patients souffrant de pathologies chroniques, comme les cancers, préoccupe les décideurs politiques. Dans ce domaine, les cliniciens ne sont pas en reste car ils souhaitent pouvoir intégrer la perception des patients, vis-à-vis de l'évolution de leur état de santé, dans le processus de décision de leur prise en charge. Les progrès réalisés à ce jour tant au niveau technologique que de celui de la connaissance de phénomènes de santé font des problématiques citées un challenge important. En effet, le cumul du progrès des connaissances, de l'évolution des technologies et l'accessibilité des outils d'applications conduisent à la grande prolifération de bases de données dans différents domaines, notamment les entreprises, l'éducation, le médical, scientifique, etc. Dans le domaine de la santé nous pouvons citer l'exemple des bases de données médico-administratives, dont l'objectif initial était la connaissance de l'activité hospitalière et qui, dans la plupart des pays occidentaux présentent de très bonnes caractéristiques d'exhaustivité, et de qualité. Nous pouvons également citer l'exemple des données issues des essais thérapeutiques. Pour les chercheurs à l'origine de ces données, une des difficultés est de pouvoir rendre les informations accessibles et compréhensibles, tout en maitrisant la gestion de volumes importants de données par une création ou une adaptation des outils existants. C'est tout l'intérêt des techniques de fouilles de données, qui s'appuient sur une stratégie très différente de celles habituellement réalisées. Au lieu de définir a priori des hypothèses et des variables d'intérêt puis de mesurer les associations statistiquement significatives entre ces variables et le phénomène à étudier, ces méthodes ne définissent pas d'hypothèses et vont rechercher dans les données toutes les informations susceptibles de nous aider dans l'analyse du problème posé et enfin, restituer les informations associées significativement au phénomène étudié.

Dans ce travail, nous nous intéressons à l'utilisation des méthodes de fouilles de données pour l'étude de l'évolution des trajectoires à partir de 2 types de bases de données. Nous montrons l'intérêt de l'application de ces méthodes au travers de deux questions d'actualité :

1 / Quelle sont les trajectoires suivies par les patients atteints de cancer et que peut-on en déduire concernant l'organisation des soins ?

2/ Quelle est l'évolution de la qualité de vie (santé perçue) chez les patients souffrant d'un cancer métastatique, et l'association entre les trajectoires de QdV et les traitements reçus, dans le cadre d'un essai thérapeutique.

Dans la première partie de cette thèse, nous présentons le travail réalisé dans le cadre d'un projet financé par l'INCA, qui traitait de la question de l'organisation des soins pour la prise en charge des patients atteints de cancers. Nous avons analysé la base médico-administrative issue du programme de médicalisation du système d'information médicale (PMSI) déployé sur tous les établissements de santé de France. La construction de trajectoires hospitalières des patients nous a permis de faire un état des lieux de la prise en charge des cancers du poumon, du colon et du sein.

Dans la deuxième partie, nous nous sommes intéressés à la santé perçue par des patients souffrant d'un cancer digestif métastatique, exprimée par des scores de qualité de vie, à partir des données recueillies au cours d'un essai thérapeutique comparatif et multicentrique. Nous présentons la stratégie mise en œuvre pour mettre en évidence des profils évolutifs de qualité de vie à partir d'un recueil de données de perception en présence de données manquantes. Ce travail a bénéficié d'une subvention du conseil régional de Bourgogne et du prix Pierre et Liliane Dusserre.

La troisième partie est consacrée à une discussion des problèmes rencontrés dans ces deux approches, suivie d'une conclusion.

CHAPITRE 1 :

ÉTUDE DES TRAJECTOIRES DE PRISE EN CHARGE DES CANCERS DANS LA RÉGION BOURGOGNE :

APPLICATION AUX CANCERS DU SEIN, DU CÔLON-RECTUM ET DU POUMON.

I. INTRODUCTION

Le projet TRAJCAN, dont l'objectif était d'étudier les trajectoires hospitalières de prise en charge des cancers, a débuté en Octobre 2010. Il s'inscrivait dans un contexte particulier lié en grande partie à la publication du nouveau plan cancer 2009-2013 intitulé « Pour un nouvel élan ». Ce dernier faisait d'une part état d'un certain nombre de difficultés persistantes, à savoir :

- La disparité dans le recours sur le territoire au diagnostic et au dépistage individuel et aux soins
- La dispersion dans l'administration des soins
- L'inégalité dans la qualité des soins

D'autre part, il annonçait la mise en œuvre prochaine du nouveau dispositif d'autorisation pour les activités de soins et de traitement du cancer.

D'une façon concomitante, l'utilisation des données de la base nationales du PMSI se démocratisait et la qualité de ses données avait beaucoup évolué. Tout ceci était propice aux développements de nouvelles méthodologies d'analyse des données dans le but de proposer des solutions aux problèmes posés.

Dans ce contexte, un état des lieux des trajectoires de soins des patients et de leurs typologies par type de cancer devait permettre d'avoir une description de l'existant (en termes de fréquentations d'établissements par les patients et de collaborations entre établissements) afin d'apporter des éléments qui pourraient être utilisés lors des campagnes d'autorisations.

II. OBJECTIF DU PROJET

Il s'agissait dans un premier temps de reconstituer les trajectoires de prise en charge des patients identifiés comme atteints de cancer à partir des données issues du PMSI. Puis, dans un deuxième temps, il convenait de décrire ces trajectoires et de construire des typologies pour une lecture et une interprétation simple des différentes situations. Ces typologies devaient pouvoir s'expliquer à partir des variables issues ou construites à partir du PMSI. Enfin, les coopérations entre les différents établissements de santé intervenant dans les différentes trajectoires devaient être étudiées de manière à identifier des associations à promouvoir. Trois localisations de cancer avaient été sélectionnées pour leurs fréquences dans la région Bourgogne : le poumon, le colon-rectum et le sein.

III. MATÉRIELS ET MÉTHODES

1- Construction des trajectoires

Il s'agissait d'une étude rétrospective, multicentrique concernant la reconstitution des trajectoires hospitalières sur le territoire national pour des patients résidant en Bourgogne. Les localisations cancéreuses étudiées étaient le poumon, le colon-rectum et le sein.

L'étude du parcours complet d'hospitalisations centré sur le patient nécessitait une approche longitudinale de séquences d'évènements de santé ordonnées dans le temps. La reconstitution de ces trajectoires reposait sur le chaînage des différents séjours hospitaliers du patient (voir annexe 1), ceci afin d'une part de repérer la première hospitalisation de chirurgie anticancéreuse et d'autre part d'assurer un suivi du patient pendant 1 an.

Les patients inclus étaient âgés d'au moins 18 ans et résidaient en Bourgogne au moment de la 1ère chirurgie contre le cancer réalisée entre 2006 et 2008, quel que soit l'établissement de santé de sa réalisation. Pour les localisations étudiées, seules les tumeurs malignes et à évolution imprévisible étaient prises en compte. Pour chaque trajectoire reconstituée une description était réalisée. Celle-ci concernait l'ensemble des établissements fréquentés, les différentes prises en charge liées ou non au cancer et le statut du patient (vivant ou décédé) au terme du dernier séjour.

Une typologie de ces trajectoires a été construite en utilisant des méthodes de classification utilisées en fouille de données. Les patients étaient regroupés selon les établissements de santé fréquentés et selon le type de prise en charge reçue, liée ou non au cancer.

Une description de cette typologie était réalisée avec les variables telles que : âge, sexe, département de résidence, les distances parcourues,…. Une description détaillée de chacune de ces variables est proposée dans l'annexe1.

2- Extraction de motifs séquentiels

L'extraction de motifs séquentiels fréquents consiste à détecter automatiquement des régularités dans des données ordonnées. Les trajectoires de soins reconstruites à partir du PMSI peuvent être vues comme des séquences d'hospitalisations, chaque hospitalisation correspondant à un ensemble de variables essentiellement qualitatives (hôpital, diagnostics, actes médicaux …).

Des algorithmes de fouille de motifs séquentiels ont été développés dès la fin des années 1990, pour traiter ce type de données, mais ils ne permettent pas de prendre en compte la

complexité des trajectoires de soins, notamment leur aspect multidimensionnel (géographique, diagnostique, thérapeutique …).

Par ailleurs, les données du PMSI sont codées à partir de terminologies possédant une structure hiérarchique et autorisent certains regroupements souvent bien utiles pour analyser et résumer l'information contenue dans les trajectoires de soins. Dans un objectif de fouille de données, une des difficultés est de choisir un niveau de granularité d'information suffisant pour extraire des résultats à la fois suffisamment spécifiques pour être intéressants et suffisamment généraux pour ne pas surcharger l'analyste lors de leur interprétation.

IV. ÉQUIPES IMPLIQUÉES DANS LE PROJET

Trois équipes de recherches étaient impliquées dans ce projet, chacun avec un objectif en relation avec son thème de recherche.

A. LE DIM DU CHU DE DIJON

Il s'agissait pour cette équipe tout d'abord d'extraire à partir des bases nationales du PMSI, les données spécifiques relatives aux patients et aux séjours de ces derniers. Puis nous avons reconstitué le suivi chronologique de ces patients avec un recul d'un an. Ces informations mises sous une forme adaptée étaient ensuite transmises aux différentes équipes. Les annexes 2 et 3 présentent respectivement le flowchart décrivant le circuit de traitement des données depuis la base nationale du PMSI jusqu'aux résultats bruts mis à disposition des équipes participantes et les résultats en termes de nombre de patients et de séjours sélectionnés pour chaque localisation dans les bases nationales. L'équipe a également réalisé une première classification des trajectoires devant servir de base de travail pour la construction de la typologie finale (voir annexe 4).

B. L'ÉQUIPE ORPAILLEUR DU LORIA À NANCY

L'équipe Orpailleur est spécialisée dans l'extraction de connaissances à partir de données. Elle a mené des travaux de classification des trajectoires reposant sur deux types d'outils : l'analyse formelle de concepts et l'extraction de motifs séquentiels fréquents. Dans un premier temps les trajectoires institutionnelles, c'est à dire la suite d'établissements fréquentés par les patients, ont été regroupées et visualisées grâce à des treillis de concepts. Ces représentations illustrent la place et l'organisation "naturelle" des établissements prenant en charge les patients de l'étude. Dans un second temps des travaux théoriques ont été menés sur le

développement de méthodes de fouille de données séquentielles. Ils ont porté en particulier sur la façon de représenter les trajectoires de soins et leur contenu, en prenant en compte l'ordre des évènements qui la composent.

C. L'ÉQUIPE DU LABORATOIRE CEREMADE DE L'UNIVERSITÉ PARIS DAUPHINE

La spécificité de cette équipe était l'analyse des données symboliques. La pré-classification réalisée par l'équipe de Dijon permettait de construire la table de données symboliques (élément sur lequel se basait toute l'analyse) en utilisant ces classes comme concept (terminologie symbolique). Ces classes étaient décrites à partir des variables soit disponibles directement dans la base nationale du PMSI (l'âge des patients, type de traitement contre le cancer, les établissements fréquentés, …), soit construites (les distances entre le lieu de résidence et l'établissement de santé, l'indice de comorbidité calculé sur chaque séjour, …). La construction d'une typologie (en 4 ou 5 classes) consistait à classifier ces concepts en utilisant la méthode des nuées dynamiques adaptée aux données symboliques. Les variables utilisées pour la classification étaient méthodiquement sélectionnées. Le résultat final était présenté sous une forme « d'histogramme » de variation.

V. RÉSULTATS

Nous présentons ci-dessous, par localisation (cf sections A, B, C) uniquement les résultats de la description des trajectoires de soins. Les résultats obtenus pour l'extraction des motifs séquentiels sont présentés dans les publications [3,4,5]. (cf section D)

A. CANCER DU POUMON

i. Description

495 patients ont été sélectionnés, ce qui correspondait à un total de 3 821 séjours d'hospitalisations. Le tableau 1 montre la répartition du nombre de patients par département de résidence et par année de sélection. La répartition par département de ces patients n'était pas statistiquement différente selon le sexe et l'âge. Le nombre d'épisodes de prise en charge observé chez ces patients était plus petit dans la Nièvre par rapport aux autres départements (p=0,064) (cf. tableau 2).

Les 94 classes obtenues après la pré-classification ont été regroupées pour construire une typologie des prises en charge en 4 classes. Le tableau 3 montre la répartition en fonction des différentes classes du nombre de patients et du nombre de trajectoires. Les 4 classes se distinguaient nettement au niveau du type de l'établissement de santé où avait eu lieu la première intervention chirurgicale contre le cancer : un CHU, un centre hospitalier (CH), un centre de lutte contre le cancer (CLCC) ou un établissement privé participant au service public hospitalier (PSPH). Sur la figure 1 on pouvait observer des barres distinctes entre les classes pour la variable premierEtab (voire annexe 1). Ainsi, nous pouvions par exemple observer que pour la classe 3, les patients résidaient majoritairement (la barre d'histogramme la plus élevée) dans le département de la Côte-d'Or au moment de l'intervention chirurgicale initiale. Cette dernière ayant eu lieu dans la plupart des cas dans une clinique privée (CL). Lorsque l'on analysait le déroulement dans son ensemble de la trajectoire de ces patients, en termes de flux entre le département de résidence, la région Bourgogne et le reste du territoire national, on constatait alors qu'ils avaient reçu des soins de proximité. En effet, la barre correspondant à la modalité « territoriale » sur la figure 1 était la plus haute dans cette classe. À l'inverse, dans la classe 2, c'est la barre « Fuite régionale » qui était la plus haute, illustrant le fait que les patients qui résidaient dans le département de l'Yonne se faisaient opérer dans des établissements de type PSPH en dehors de la région Bourgogne. La figure 3 correspondant à ce département montrait qu'il s'agissait principalement de la clinique Marie Lannelongue en région parisienne. Lorsqu'il s'agissait des prises en charge répétitives après la

chirurgie (les séances de chimiothérapie), les patients de la classe 3 poursuivaient les soins dans les cliniques privées alors que ceux de la classe 2 se déroulaient dans les centres hospitaliers de proximité (voir la variable hôpitaux sur la figure 1).

La figure 3 montre l'organisation des établissements qui prenaient en charge les patients atteints d'un cancer du poumon. Elle est obtenue à partir d'un treillis formé de 302 concepts dont on a retenu les plus fréquents (effectifs > 5).

Au centre de la figure en haut apparait le CHU de Dijon qui a accueilli 146 patients distincts. A sa gauche, le CH d'Autun a hospitalisé 11 patients. Entre les deux, un nœud montre que 8 patients étaient communs à ces établissements. Ainsi, chaque nœud (ou concept) sur la figure représente à la fois un groupe de patients et l'ensemble des établissements qu'ils ont fréquentés. Cet ensemble peut être composé de 1 (symbole hôpital), 2 (symbole nuage) voire 3 (symbole losange) hôpitaux. Certaines icones (double cercle) montrent que tous les patients d'un hôpital ont fréquenté l'autre : par exemple, les 5 patients hospitalisés en SSR au CH de Chatillon étaient aussi connus du CHU de Dijon. La topologie de ce graphe montre à la fois l'organisation essentiellement géographique des trajectoires de soins mais aussi la place prise par certains établissements. Le CHU de Dijon est par exemple un nœud par lequel transitaient beaucoup de trajectoires quantitativement (146 patients) mais aussi qualitativement puisqu'il partageait des patients avec de nombreux autres établissements (au moins 8). Plusieurs centres accueillaient les fuites régionales (APHP, Marie Lannelongue, Hospices civils de Lyon) mais très souvent en coopération avec un centre Bourguignon : par exemple, 31 des 79 patients vus par l'APHP ont été également hospitalisés au CH de Nevers. C'est d'ailleurs la quasi-totalité des patients (31/34) vus par le CH de Nevers, ce qui suggère une forte interaction entre ces deux établissements.

ii. Résultats géographiques pour le cancer du poumon

Les résultats en termes de modalités de fréquentation du système de soins divergent fortement entre le cancer du poumon et les autres localisations, ce qui n'est pas surprenant eu égard aux protocoles de prise en charge chirurgicale, nettement plus lourds dans le cas du poumon. Le travail sur le cancer du poumon met en lumière des trajectoires davantage tournées vers les grands centres urbains qui concentrent d'importants centres hospitaliers. Seuls les patients de Saône-et-Loire privilégient les petites structures. Notons que ces résultats s'appuient sur les bases du PMSI 2006 à 2008, date de mise en œuvre des seuils minimaux d'activité en chirurgie carcinologique suite aux recommandations de l'InCa. Le système général des

trajectoires de prise en charge chirurgicale du cancer du poumon en région Bourgogne suit donc les trois grandes interfaces qui modèlent les déplacements dans la région : régions parisienne et lyonnaise au nord et au sud, Dijon et sa périphérie, qui recrutent des patients originaires de toute la région. Le traitement du cancer du poumon oriente donc les patients vers des déplacements longs, qui privilégient les structures exploitant des plateaux chirurgicaux importants (cf. Figure 2).

iii. Tableaux et figures

Tableau 1: Répartition du nombre de patients sélectionnés pour la localisation pulmonaire par département de résidence et par année de réalisation de la première chirurgie

Année\Département	**Côte-d'Or (21)**	**Nièvre (58)**	**Saône-et-Loire (71)**	**Yonne (89)**	**Total**
2006	14	23	37	13	**87**
2007	48	35	73	46	**202**
2008	49	27	77	53	**206**
Total	**111**	**85**	**187**	**112**	**495**

Tableau 2: Caractéristiques des patients pour la localisation pulmonaire

Variables		Côte d'Or	Nièvre	Saône-et-Loire	Yonne	Total	p*
Effectif		111	85	187	112	**495**	
Sexe	**Hommes**	85	63	150	95	393	0,253
	Femmes	26	22	37	17	102	
Age (*Moyenne±écart type)*		63±10	63±10	64±10	64±10	64±10	0,404
Nb d'épisodes		8±5	6±6	8±6	9±10	8±10	0,064

* *significativité*

Tableau 3: Description des classes en nombre de trajectoires et de patients pour la localisation pulmonaire

Typologie	Classe 1	Classe 2	Classe 3	Classe 4	Total
Nb de trajectoires (%)	15 (16,0)	24 (25,5)	19 (20,2)	36 (38,3)	94
Nb de patients (%)	42 (8,5)	68 (13,7)	138 (27,8)	247 (50,0)	495

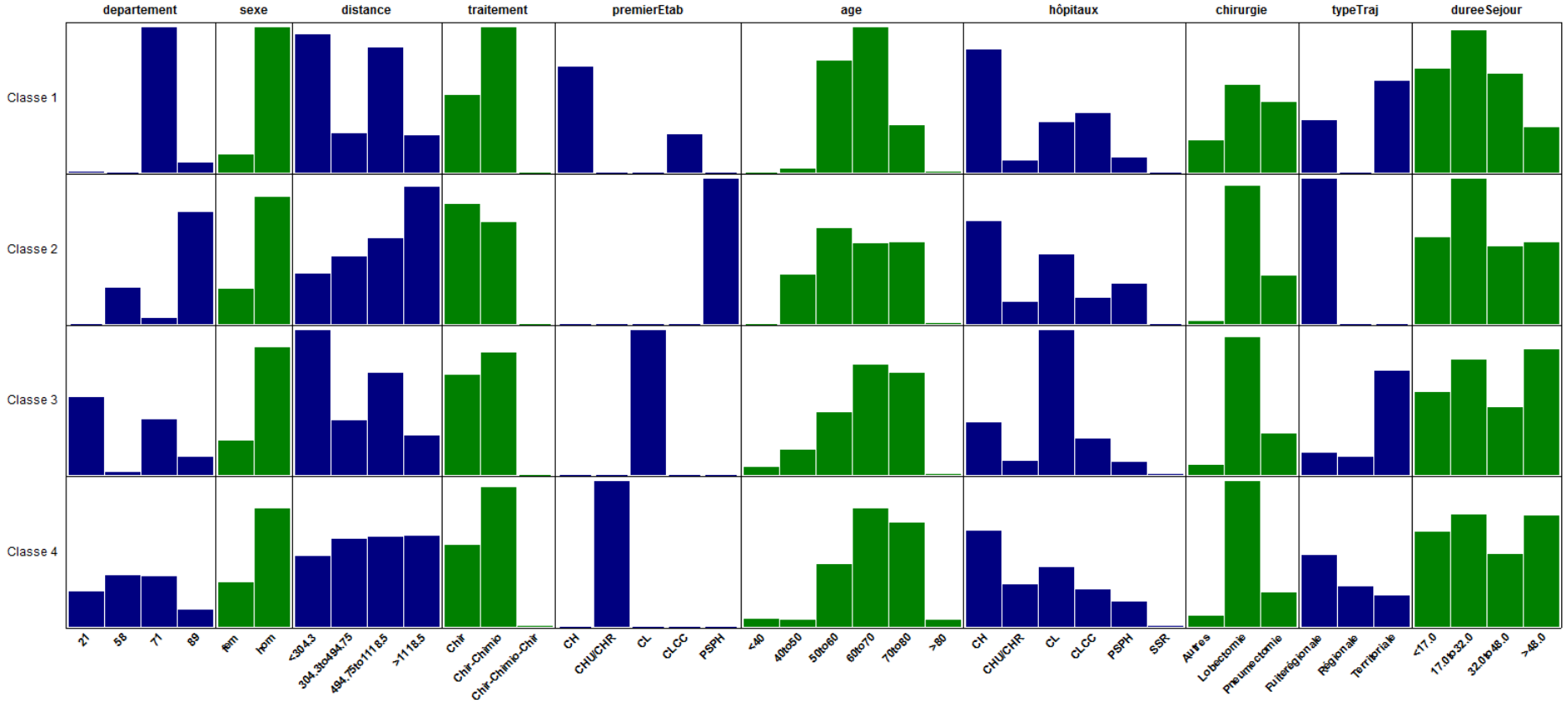
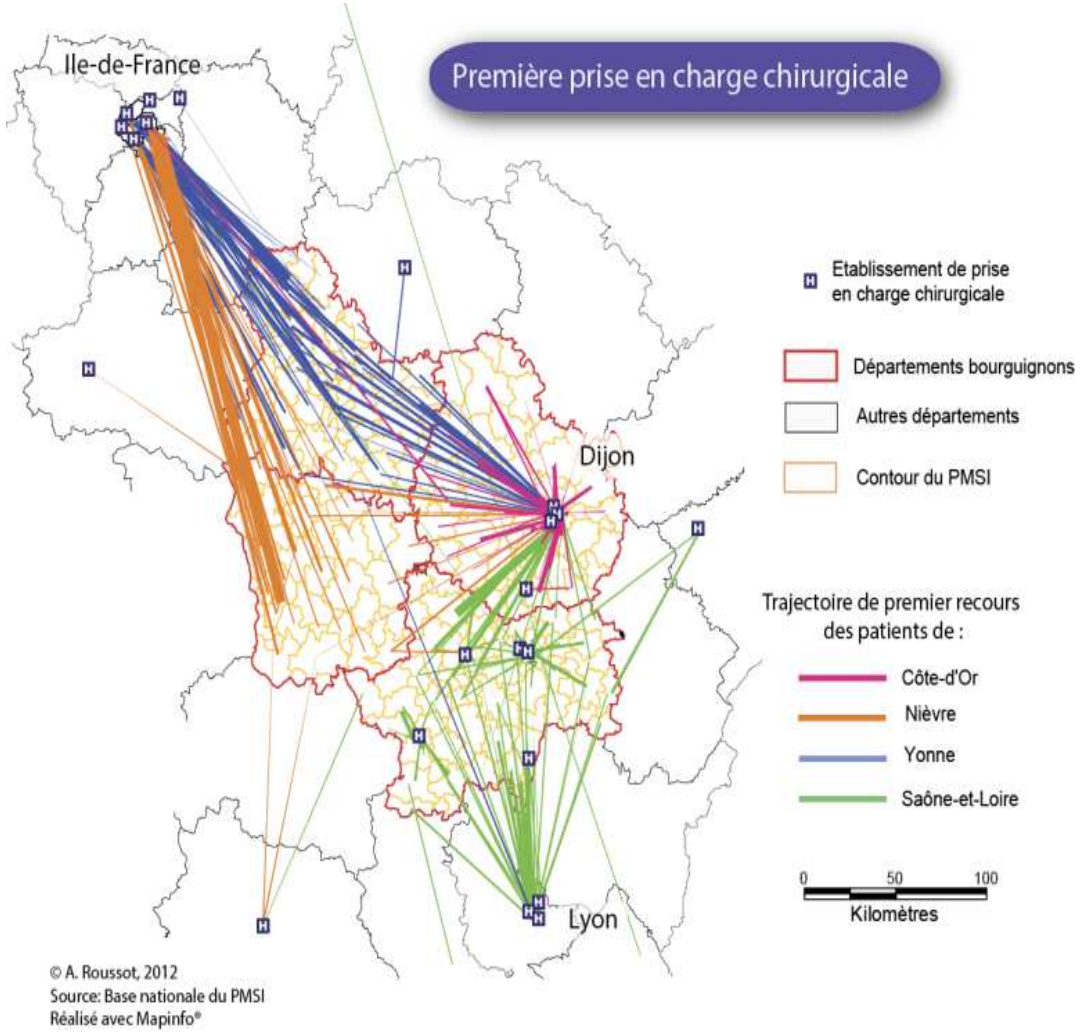


Figure 1: Typologie en 4 classes des trajectoires de prises en charge du cancer du poumon des patients résidant dans la région Bourgogne



Pour la Côte-d'Or :
CHU Dijon : **55** patients
Clinique de Chenôve : **11** patients
Clinique Benigne Joly : **6** patients

Pour l'Yonne :
Clinique Marie Lannelongue : **30** patients
Hôpital Foch : **18** patients
AP-HP : **15** patients
CHU Dijon : **27** patients

Figure 2:Analyse spatiale des flux de patients résidant en Côte d'Or et dans l'Yonne

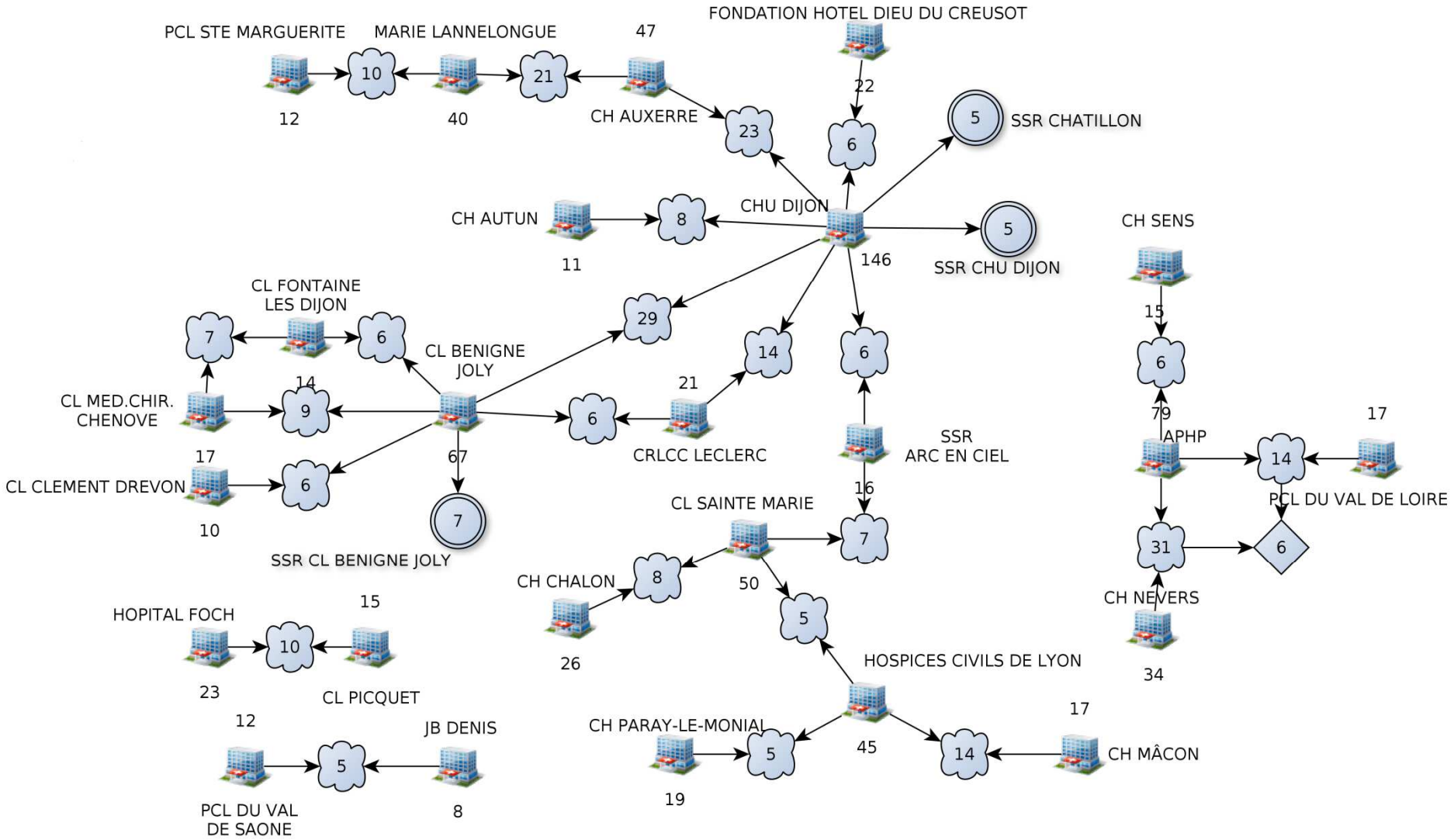


Figure 3 : treillis de trajectoires institutionnelles de patients atteints de cancer du poumon

B. CANCER COLO-RECTAL

iv. Description

2 639 patients ont été sélectionnés, ce qui correspondait à un total de 27 468 séjours d'hospitalisation. Le tableau 4 montre la répartition du nombre de patients par département de résidence et par année de sélection. La répartition par département de ces patients n'était pas statistiquement différente selon le sexe et l'âge. Le nombre d'épisodes de prise en charge était significativement plus petit chez les patients résidant en Côte d'Or par rapport aux autres départements (p=0,033) (cf. tableau 5).

Les 150 classes obtenues après la pré-classification ont été regroupées pour construire une typologie des prises en charge en 5 classes. Le tableau 6 montre la répartition en fonction des différentes classes du nombre de patients et du nombre de trajectoires. Cette typologie est essentiellement caractérisée par une prise en charge de proximité lisible sur l'ensemble des 5 classes. En effet, sur la figure 4 la modalité territoriale avait la barre de fréquence la plus élevée pour l'ensemble de la typologie. De plus le type d'établissement fréquenté par les patients pour les séances de chimiothérapie ou d'autres traitements complémentaires était pratiquement le même que celui où avait été réalisé la première intervention chirurgicale. Par exemple, les patients de la classe 4 qui résidaient majoritairement en Côte-d'Or poursuivaient leur prise en charge dans un CHU et très probablement celui de Dijon.

La figure 6 montre le treillis de concepts obtenu pour les trajectoires de patients soignés pour cancer du côlon ou du rectum. Sur cette figure, la taille des concepts est proportionnelle au nombre de patients qui le composent. Quatre centres accueillaient plus de 200 patients : le CHU de Dijon, le CH d'Auxerre, la Clinique Sainte Marie à Chalon/Saône et la Clinique Drevon à Dijon. Les interactions concernant le plus grand nombre de patients sont relevées entre la Clinique Drevon et la Clinique Joly, et entre la Clinique Sainte Marie (>50 patients communs). Comme pour le cancer du poumon, les filières de prise en charge suivaient une logique géographique entre établissements d'une même zone : Dijon, Auxerre, Chalon/Saône, Nevers, Macon. Les échanges avec des hôpitaux hors Bourgogne semblaient moins importants que pour le cancer du poumon, hormis pour les établissements de l'Yonne avec l'APHP.

Sur cette figure apparaissent également des établissements ayant peu de patients communs avec d'autres (<15), souvent parce qu'ils en traitaient peu eux-mêmes. Néanmoins, certains d'entre eux pouvaient en prendre en charge un nombre important. L'exploration par treillis de

concepts permet d'examiner les interactions en détail comme dans le tableau 7, qui montre toutes les trajectoires incluant le CH de Beaune concernant au moins 5 patients.

v. Résultats géographiques pour le cancer colorectal

A l'inverse des orientations générales observées pour les patients atteints de cancer du poumon, les résultats sur le traitement du cancer colorectal montrent des trajectoires plus courtes, tournées vers des structures de proximité directe. Les patients, plus nombreux que ceux atteints de cancer primitif du poumon, orientent leurs déplacements vers les centres de soins les plus proches de leur territoire de résidence, même si des trajectoires plus étendues sont toujours observables, vers Paris ou Lyon. La dichotomie entre les résultats pour les deux localisations cancéreuses s'explique par l'offre de soins disponible et la présence de spécialistes dans les centres de soins concernés, par la plus forte incidence des cas de cancers colorectaux, plus facilement pris en charge, et enfin par les liens forts qui existent entre certains territoires bourguignons et les métropoles avec lesquelles ils entretiennent une histoire et des échanges multiséculaires, comme entre la Nièvre et Paris (cf. Figure 5).

vi. Tableaux et figures

Tableau 4: Répartition du nombre de patients sélectionnés pour la localisation colo-rectal par département de résidence et par année de réalisation de la première chirurgie

Année\Département	**Côte-d'Or (21)**	**Nièvre (58)**	**Saône-et-Loire (71)**	**Yonne (89)**	**Total**
2006	168	131	303	154	**756**
2007	149	149	338	189	**913**
2008	438	158	341	215	**970**
Total	**661**	**438**	**982**	**558**	**2 639**

Tableau 5: Caractéristiques des patients pour la localisation pulmonaire

Variables	Côte d'Or	Nièvre	Saône-et-Loire	Yonne	p*
Effectif	661	438	982	558	
Sexe (% femme)	46	40	43	42	0,160
Age (*Moyenne±écart type)*	71±13	71±11	71±12	70±12	0,436
Nb d'épisodes	9±9	11±11	11±12	11±11	0,033

* *significativité*

Tableau 6: Description des classes en nombre de trajectoires et de patients pour la localisation colo-rectale

Typologie	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total
Nb de trajectoires (%)	32 (21,3)	21 (14)	31 (20,7)	36 (24,3)	30 (20,0)	150
Nb de patients (%)	194 (7,4)	529 (20)	977 (37,0)	157 (6,0)	782 (29,6)	2 639

Tableau 7: Trajectoires incluant le CH de Beaune pour la localisation colo-rectale (d'effectif > 5).

Concept	Effectifs
CENTRE HOSPITALIER DE BEAUNE	94
CENTRE HOSPITALIER DE BEAUNE, C.H.U. DE DIJON	11
CENTRE HOSPITALIER DE BEAUNE, CLINIQUE CLEMENT DREVON	9
CENTRE HOSPITALIER DE BEAUNE, CLINIQUE MEDICO-CHIRURGICALE	5
MAISON DE REPOS LA FOUGERE, CENTRE HOSPITALIER DE BEAUNE	5
CENTRE HOSPITALIER DE BEAUNE, CLINIQUE DE FONTAINE	5

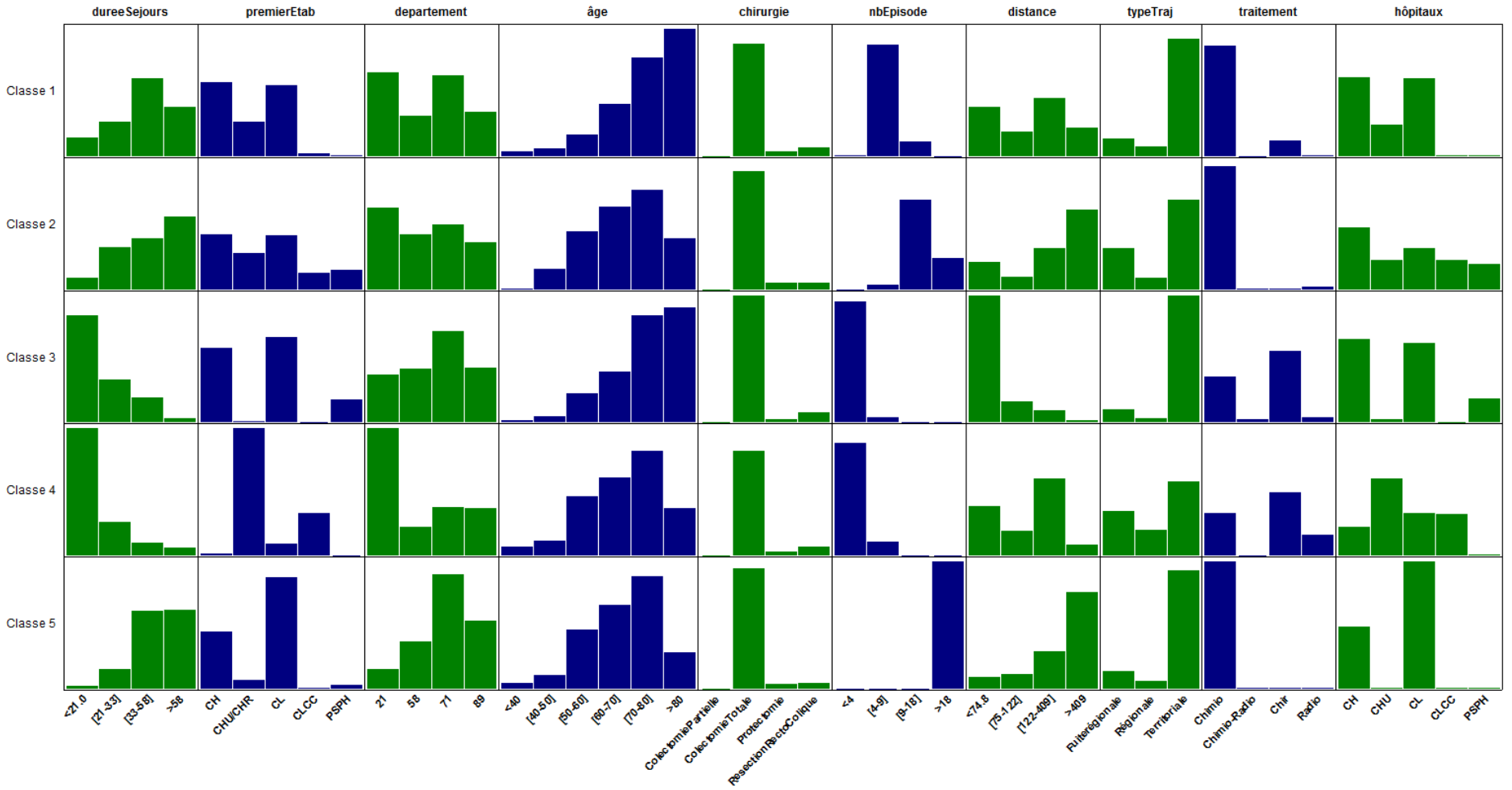


Figure 4 : Typologie en 5 classes des trajectoires de prises en charge du cancer colo-rectal des patients résidant dans la région Bourgogne

Trajectoires de chirurgie colorectale, première prise en charge des patients résidant dans la Nièvre et en Saône-et-Loire de 2006 à 2009

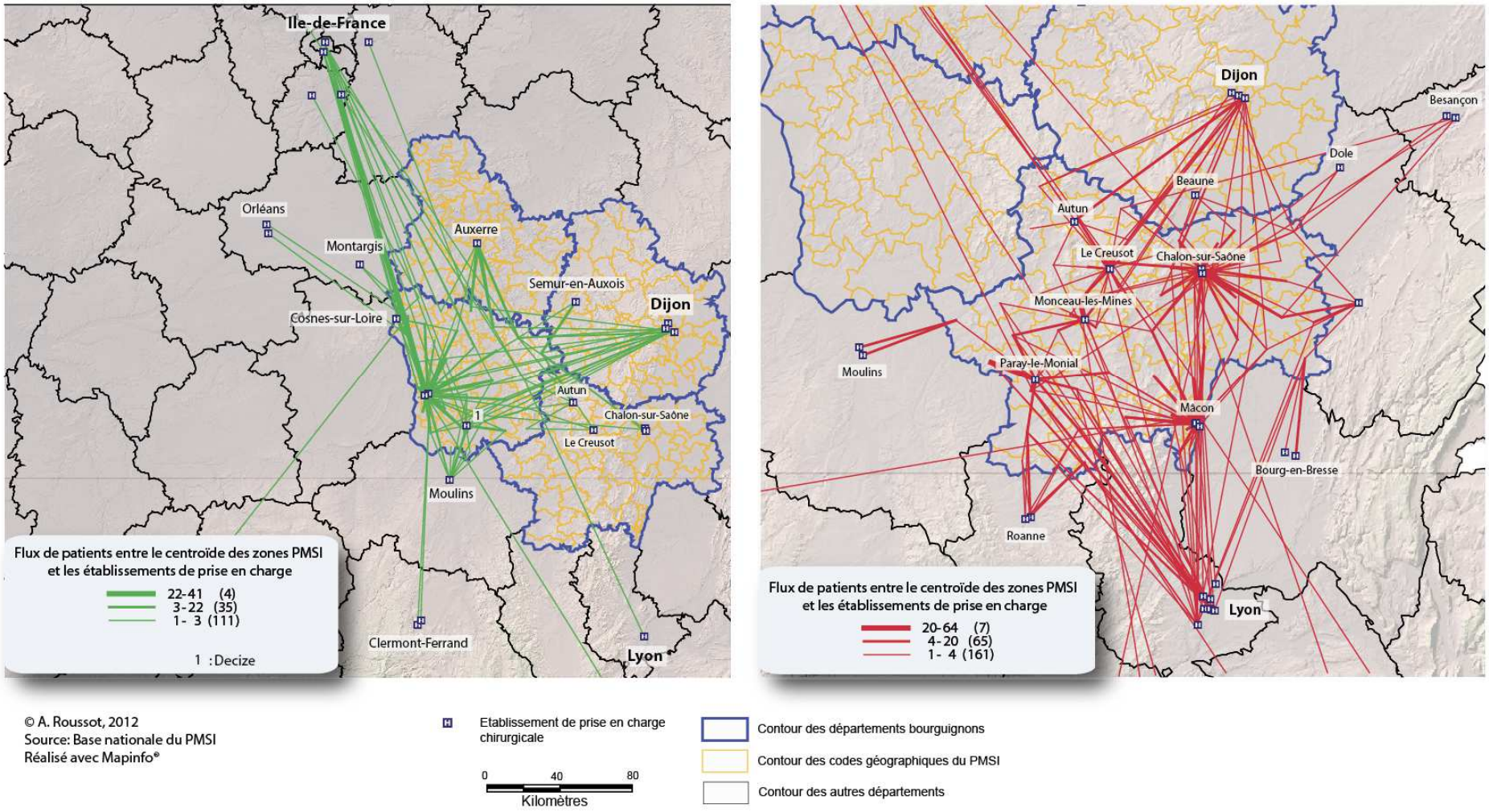


Figure 5: Trajectoire de premier recours de chirurgie pour les patients de la Nièvre et de la Saône-et-Loire

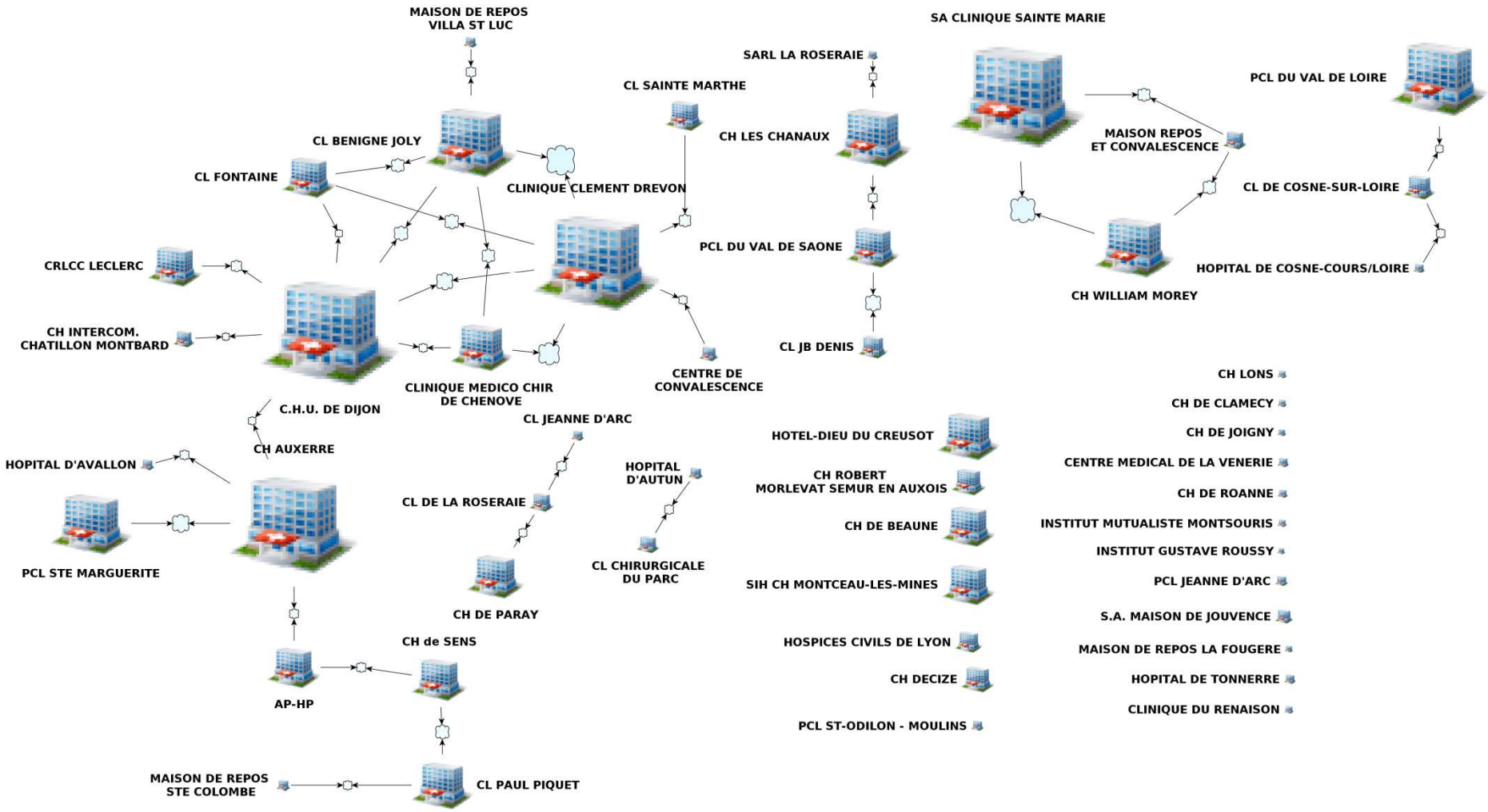


Figure 6 : treillis de trajectoires institutionnelles de patients atteints de cancer du côlon-rectum

C. CANCER DU SEIN

vii. Description

4 441 patientes ont été sélectionnées, ce qui correspondait à un total de 52 441 séjours d'hospitalisation. Le tableau 7 montre la répartition du nombre de patientes par département de résidence et par année de sélection. La répartition par département de ces patientes n'était pas statistiquement différente selon le sexe. Le nombre d'épisodes de prise en charge était significativement doublé pour les patientes résidant en Côte d'Or par rapport aux autres départements ($p<10-3$) (cf. tableau 8).

Les 150 classes obtenues après la pré-classification ont été regroupées pour construire une typologie des prises en charge en 4 classes. Le tableau 9 montre la répartition en fonction des différentes classes du nombre de patients et du nombre de trajectoires. Comme pour la localisation pulmonaire, les classes de trajectoire pour le sein sont en grande partie déterminées à la fois par le lieu de réalisation de l'intervention chirurgicale initiale et par le flux spatial des patientes. Par exemple, la classe 2 sur la figure 7 rassemblait les patientes qui résidaient majoritairement dans le département de la Saône-et-Loire et qui avaient été opérées dans une clinique privée. Pour ces patientes, l'ensemble de la prise en charge s'était déroulé à l'intérieur du département de résidence (la modalité territoriale était la plus fréquente). De plus, la poursuite des soins après la chirurgie se déroulait également dans les cliniques privées, comme on le voit sur la figure 7 où la modalité CL était majoritairement représentée pour les variables premierEtab et hôpitaux.

La figure 8 montre le treillis correspondant aux trajectoires de soins des patientes atteintes de cancer du sein. On remarque l'importance prise par un seul centre qui traitait près d'un quart des patientes de la région pour cette localisation. Une filière importante se constituait par ailleurs dans le sud de la région autour des établissements de Chalon et Macon. Quelques établissements accueillaient plus de 200 patients, mais les interactions entre hôpitaux étaient moins présentes que pour les localisations pulmonaire et du colon-rectum.

viii. Tableaux et figures

Tableau 8: Répartition du nombre de patients sélectionnés pour la localisation du sein par département de résidence et par année de réalisation de la première chirurgie

Année\Département	**Côte-d'Or (21)**	**Nièvre (58)**	**Saône-et-Loire (71)**	**Yonne (89)**	**Total**
2006	469	198	488	271	**1 426**
2007	458	184	559	336	**1 537**
2008	443	184	561	290	**1 478**
Total	**1 370**	**566**	**1 608**	**897**	**4 441**

Tableau 9: Caractéristiques des patients pour la localisation du sein

Variables	Côte d'Or	Nièvre	Saône-et-Loire	Yonne	p*
Effectif	1 370	566	1 608	897	
Sexe (% femme)	98,8	99,5	99,1	99,2	0,561
Age (*Moyenne±écart type)*	61±13	62±13	63±14	62±13	0,003
Nb d'épisodes	19±16	8±11	9±11	9±11	$<10^{-3}$

* *significativité*

Tableau 10: Description des classes en nombre de trajectoires et de patients pour la localisation du sein

Typologie	Classe 1	Classe 2	Classe 3	Classe 4	Total
Nb de trajectoires (%)	46 (18,4)	57 (22,8)	109 (43,6)	38 (15,2)	250
Nb de patients (%)	1 401 (31,6)	1 786 (40,2)	583 (13,1)	671 (15,1)	4 441

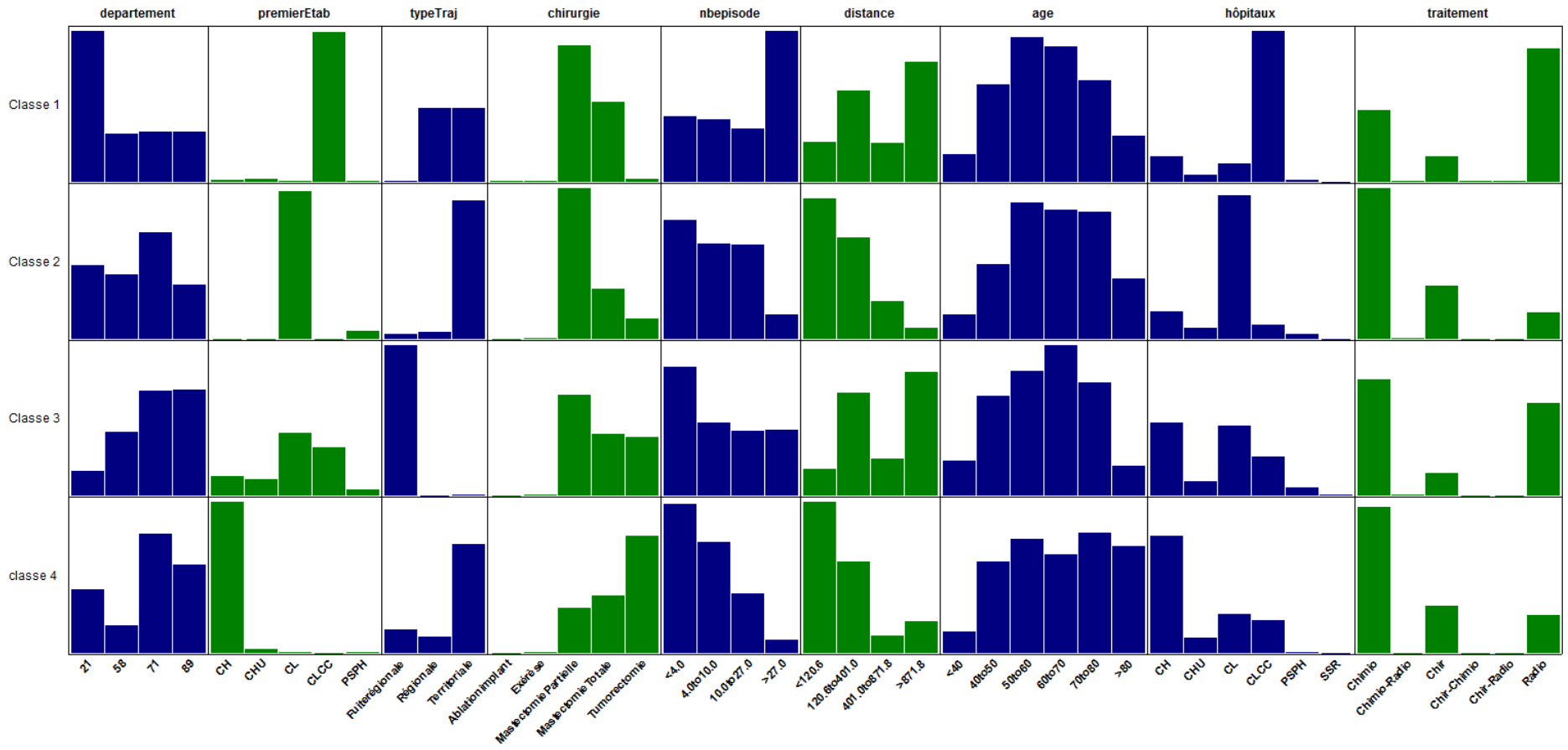


Figure 7: Typologie en 4 classes des trajectoires de prises en charge du cancer du sein des patients résidant dans la région Bourgogne

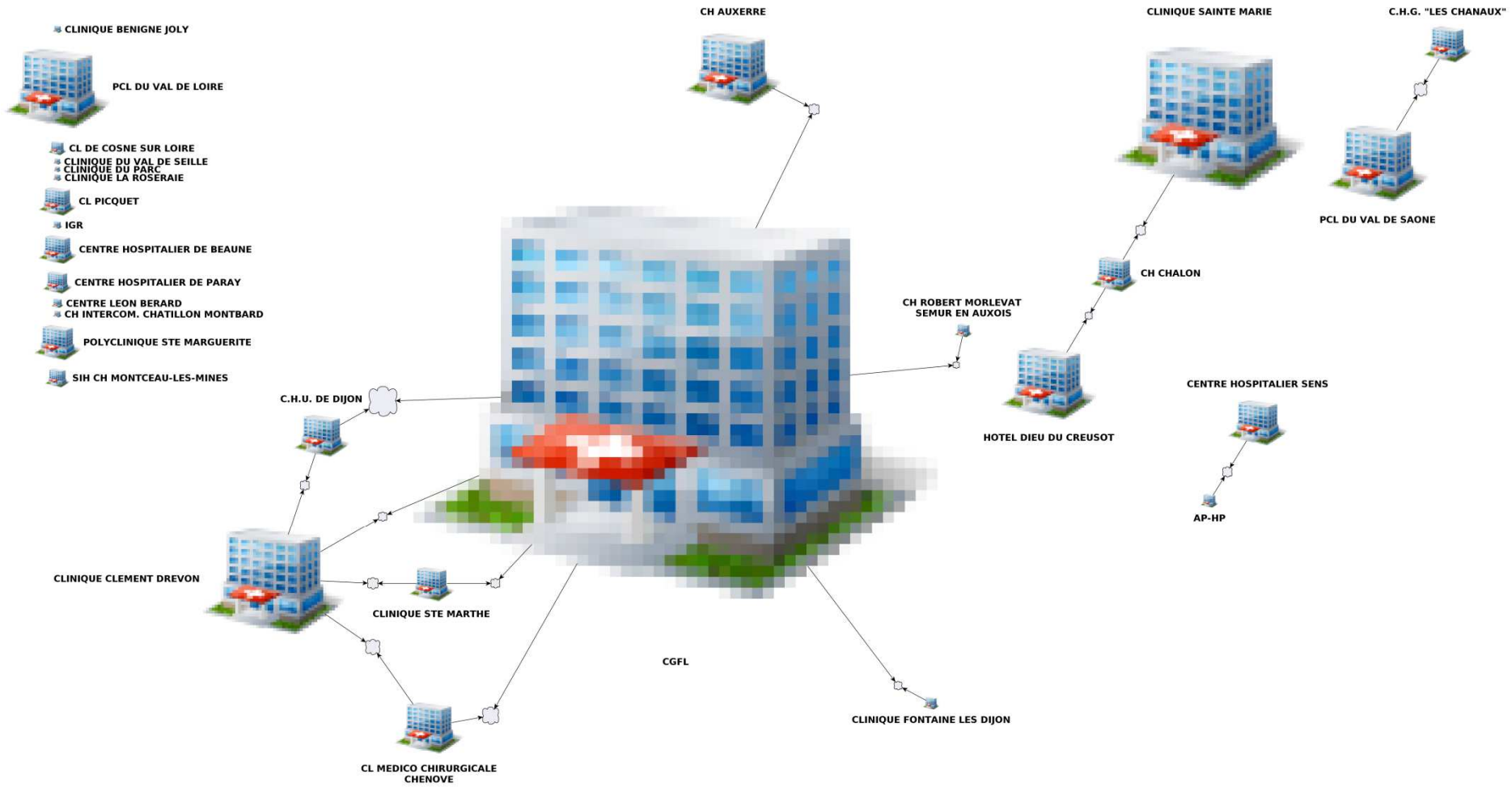


Figure 8 : treillis de trajectoires institutionnelles de patients atteints de cancer du sein

ix. 3. Analyse supplémentaire sur les coûts des trajectoires

Il nous a semblé intéressant de compléter la description des trajectoires par une analyse des coûts. Nous avons élaboré une méthode de classification des trajectoires basées sur les co-morbidités au court d'un an de suivi dans les hospitalisations des patientes atteintes d'un cancer du sein. Nous avons étendu l'étude à la France entière, pour l'année 2009, afin de disposer d'effectif suffisant des trajectoires et d'avoir une vision globale des coûts de la prise en charge hospitalière du cancer du sein en France. Nous nous sommes intéressés au coût cumulé des hospitalisations sur une année par patiente et par trajectoire.

Cette étude qui a porté sur 57 552 patientes, a permis d'isoler 19 trajectoires (cf. article publié en 2013 dans medical informatics and decision making), avec un coût moyen par trajectoire de 1600 euros.

Bien évidemment les co-morbidités les plus sévères sont associées aux coûts les plus élevés (cf. tableau ci-dessous issu de l'article). Les coûts les plus faibles sont observés pour les patients qui avaient un carcinome in situ (6957 euros) et la présence de soins palliatifs est associée aux coûts les plus élevés (26139 euros).

Cette étude a montré l'intérêt de l'utilisation des méthodes de classification automatiques, pour analyser les trajectoires de soins à partir des bases de données médico-administratives.

Table 4 Statistics by concepts. Statistics are computed on a per patient basis

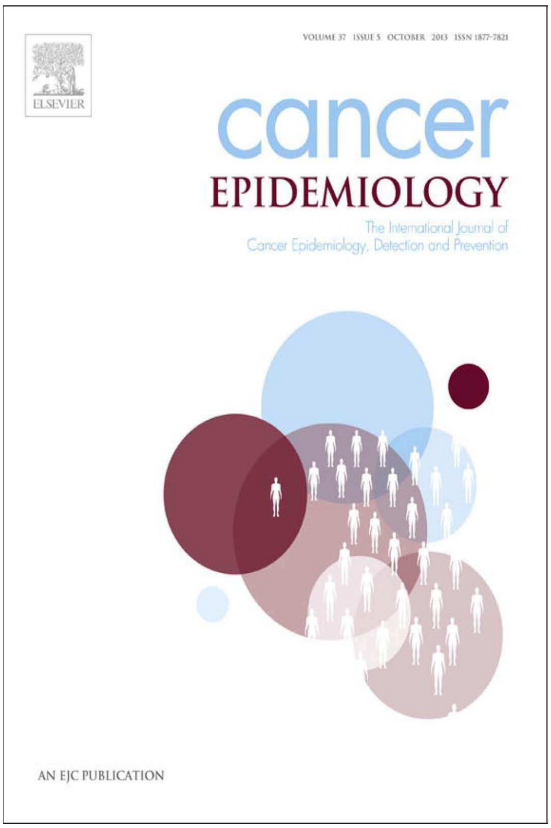
Intent	Patients	Cost	Stays		Chemotherapy sessions		Death rate	Age
	n	€	n	Cum. length	n	Cost	%	
D05	5034	6957	1.9	5.9	0.6	669	0	57.6
C50,H25	517	9499	3.1	8.2	1.1	1165	1	75.9
∅	**57552**	**9600**	**2.0**	**7.3**	**2.9**	**3090**	**1**	**60.4**
C50	53535	9902	2.0	7.5	3.1	3318	1	60.6
D05,Z421	482	11471	3.1	11.8	0.3	306	0	50.4
C50,D05	1017	12384	2.8	9.5	3.0	3109	1	56.0
C50,Z421	1339	13484	3.2	12.3	2.0	2055	0	51.2
C50,N61	445	15362	3.5	15.3	3.5	3737	1	60.7
C50,Z452,Z511	13214	15736	2.9	7.8	7.9	8341	1	56.1
C50,Z511	20820	16113	2.7	8.4	8.1	8531	1	55.3
C50,C77	863	16590	3.0	9.6	5.9	6266	1	57.6
C50,C77,Z452,Z511	348	17803	3.5	9.4	7.4	7822	1	57.0
C50,D05,Z511	350	18039	3.1	9.7	8.6	9034	1	53.4
C50,C77,Z511	687	18351	3.1	9.6	7.5	7871	1	55.7
C50,Z421,Z511	332	19946	3.6	12.4	7.9	8288	1	48.5
C50,D61,Z452,Z511	420	22319	4.5	15.5	8.1	8531	3	57.4
C50,D61,Z511	622	22598	4.4	16.4	8.0	8396	3	56.6
C50,C79	372	23052	4.0	28.6	6.3	6587	24	58.9
C50,Z515	365	26139	4.5	43.2	4.2	4371	69	65.5

D. TRAVAUX SUR L'EXTRACTION DE MOTIFS SÉQUENTIELS

Nous avons élaboré l'algorithme MMISP (Mining Multidimensional Itemsets Sequential Patterns), une méthode destinée à l'exploration des données séquentielles que sont les trajectoires de soins, en prenant en compte leurs aspects multidimensionnels et multiniveaux de granularité [3, 4, 5]. Cet algorithme permet de prendre en compte l'ordre des hospitalisations, la structure hiérarchique de la classification internationale des maladies (CIM 10) ou d'autres classifications pour caractériser de manière automatique les trajectoires de soins les plus fréquentes. Ainsi par exemple, si une première analyse conduit à définir 2 trajectoires peu fréquentes, l'algorithme MMISP va permettre d'étudier l'intérêt de leur regroupement, en généralisant leur description grâce à la structure hiérarchique de la CIM 10 : deux trajectoires reposant sur des comorbidités cardiovasculaires peuvent ainsi être regroupées.

VI. PUBLICATIONS

- G. Nuemi, F. Afonso, A. Roussot, L. Billard, J. Cottenet, E. Combier, E. Diday, C. Quantin : classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy, Cancer Epidemiol. 2013;37(5):688-96.
- A. Roussot, E. Combier, G. Nuemi, J.M. Amat-Roze, C. Quantin, Analyse spatiale des trajectoires de prise en charge des patients atteints de cancer primitif du poumon en région Bourgogne, Journal d'Économie Médicale 2012, Vol. 30, n° 2
- Elias Egho, Nicolas Jay, Chedy Raïssi, Gilles Nuemi, Catherine Quantin, Amedeo Napoli: An Approach for Mining Care Trajectories for Chronic Diseases. Artificial Intelligence in Medicine 2013. Lecture Notes in Computer Science Volume 7885, 2013, pp 258-267: 258-267
- Jay N, Nuemi G, Gadreau M, Quantin C: A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer.BMC Med Inform Decis Mak. 2013, 13:130.



Cancer Epidemiology 37 (2013) 688–696

Contents lists available at SciVerse ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Classification of hospital pathways in the management of cancer: Application to lung cancer in the region of burgundy

G. Nuemi [a,d], F. Afonso [b], A. Roussot [a], L. Billard [e], J. Cottenet [a], E. Combier [a], E. Diday [c], C. Quantin [a,d,*]

[a] Service de Biostatistique et d'Information Médicale, Centre Hospitalier Universitaire, 21000 Dijon, Boulevard Jeanne d'Arc BP 77908, 21079 Dijon Cedex, France
[b] Syrokko, Paris, France
[c] CEREMADE CNRS UMR 7534, Université de Paris, Dauphine, 75775 Paris Cedex 16, France
[d] INSERM, U866, Université de Bourgogne, 21000 Dijon, France
[e] Department of Statistics, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 17 August 2012
Received in revised form 16 June 2013
Accepted 19 June 2013
Available online 10 July 2013

Keywords:
Lung neoplasm
Hospital information systems
Epidemiology
Medical record linkage
Management care pathways
Clustering

ABSTRACT

Context: The evaluation of national cancer plans is an important aspect of their implementation. For this evaluation, the principal actors in the field (doctors, nurses, etc.) as well as decision-makers must have access to information that is reliable, synthetic and easy to interpret, and which reflects the implementation process in the field. We propose here a methodology to make this type of information available in the context of reducing inequalities with regard to access to healthcare for patients with lung cancer in the region of Burgundy. *Methods*: We used the national medico-administrative DRG-type database, which gathers together all hospital stays. By using this database, it was possible to identify and reconstruct the care management history of these patients. That is, by linking together all attended hospitals, sorted chronologically. Eligible patients were at least 18 years old, whatever the gender and had undergone surgery for their lung cancer. They had to be residents of Burgundy at the time of the first operation between 2006 and 2008. Patient's pathway was defined as the sequence of all attended hospitals (hospital stays) during the year of follow up linked together using an anonymised patient identifier. We then constructed a pathway typology of pathway using an unsupervised clustering method, and conducted a spatial analysis of this typology. *Results*: Between 2006 and 2008, we selected 495 patients in the 4 administrative departments of the Burgundy region. They accounted for a total of 3821 stays during the year of follow-up. There were 393 men (79%) and the mean age was 64 (95% confidence interval: 63–65) years. We reconstructed 94 pathways (about five per patient). Here, neighbourhood's cares accounted for 41% of them, while 44% included a surgical intervention outside the region of Burgundy. We constructed a pathway typology with five classes. Spatial analysis showed that the vast majority of initial surgeries took place in the major regional centres. *Conclusion*: The construction of a pathway typology leads to better understanding of the reasoning that lies behind the movements of patients. It opens the way for analysis of the collaboration between the different healthcares establishments attended, which should bring to light associations that need to be developed.

1. Introduction

In developed countries, the management of severe, chronic disease, such as cancer, is a major concern in public health given its growing prevalence, and requires specific, optimal organisation of the healthcare system. Public health policies aim to guarantee the best care management for everyone, and are based on a set of recommendations made by experts. This constitutes a sort of (national) references guide, which in certain countries is called the "cancer plan" [1–4]. Evaluation of the implementation of these policies is an essential step in the "virtuous circle" for improved quality, as described by the "Deming wheel" [5,6]. This evaluation is very often based on longitudinal studies, with national scope whenever possible, the aim of which is to take stock of care management. Such studies require data to be collected as close as possible to the participants in the healthcare system, both patients and healthcare professionals.

Today, most developed countries have upgraded their health-care information technology systems (HITS) to systematically and regularly record medico-administrative data on hospital care management (HCM) [7,8]. The HCM is classified according to medical and economic homogeneity, using a method inspired by

* Corresponding author at: Service de Biostatistique et d'Information Médicale, Centre Hospitalier Universitaire, 21000 Dijon, Boulevard Jeanne d'Arc BP 77908, 21079 Dijon Cedex, France.
E-mail address: catherine.quantin@chu-dijon.fr (C. Quantin).

1877-7821/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.canep.2013.06.007

Diagnosis Related Groups (DRG) developed in the late seventies by Professor Robert Fetter and his team (Yale University, United States) [9,10]. In certain cases, these databases allow the linkage of hospital stays for individual patients, and thus a longitudinal analysis of their care management. With hospital stays sorted chronologically, the linkage is done using an anonymised patient identifier that guarantees anonymity [10–15]. The chain resulting is called patient's pathway or individual pathway. In France, these databases are now accessible to researchers, which allow them to undertake longitudinal epidemiological studies [10,11,16] notably for cancer [17,18].

Of course, these longitudinal studies present the usual difficulties related not only to the analysis of repeated measures with missing data, but also to the need to take into account the spatial–temporal dimension of care management as well as its multidisciplinary aspect. In order to take this into account, it is possible to construct "individual pathway's profiles" or "pathway typology" before any statistical modelling, and then to classify each patient to one of these pathway's profiles. The description of care management experienced by a group of patients could thus be summarised with a description of the corresponding items of the pathway typology, thus facilitating the interpretation made by policy decision-makers and by healthcare professionals.

The aim of this work was to propose a method to construct these pathway's profiles by using data-mining techniques [19,20]. From an example, lung cancer, we will show that it is possible to construct a pathway typology using variables that are defined at the patient level and are commonly available in medico-economic databases. We also propose a spatial representation of these pathway's profiles, which will lead to clearer understanding of the dynamics of patient's movements.

2. Materials and methods

2.1. Materials and population

This is a retrospective multicentre study concerning the reconstitution and classification of hospital care management pathways. This study concerns patients with primary lung cancer living in the region of Burgundy France, and with surgery as the first treatment for their cancer between 2006 and 2008.

We worked on national medico-administrative data from the "Programme de médicalisation du système d'information (PMSI)". These data correspond to anonymous hospitalisation abstracts, which were collected, as required by law, in healthcare establishments between 2004 and 2009 in a standardised form. They describe hospital stays in classical medical units, in follow-up care units or in structures for hospitalisation at home. In order to link data, a unique anonymous number is attributed to every patient and included in the hospitalisation abstracts. The process used to generate these numbers guarantees the confidentiality of personal information. The administrative part of the abstracts consists of the patient's individual characteristics (age, gender, place of residence), information relative to the hospital stays (duration, type of hospitalisation), as well as the establishments attended (identification number, category). The medical part essentially comprises the diagnostic codes according to the International Classification of Diseases 10th revision (ICD 10) of the World Health Organisation (WHO) and medical acts coded according to a common, standardised nomenclature.

In this study, lung cancer designates primary bronchial cancers. Classically, this family of cancers is divided histologically into two groups: non-small cell lung carcinoma (NSCLC), which accounts for 80–90% of cases and small-cell lung carcinoma (SCLC) [21,22]. Seven stages of development can be distinguished for NSCLC [23] from the least advanced (stage IA) to the most advanced (stage IV). Around 20% of patients are classified stage I/II, 20–30% stage IIIA or IIIB, and the rest stage IV. For these tumours, surgery is the principal management strategy, and is the best treatment for stages I, II and IIIA. Chemotherapy comes in second place and can be administered either before the surgery (neo-adjuvant chemotherapy), or 4–8 weeks after the surgery (adjuvant chemotherapy), or even as the initial treatment for advanced cases (stages IIIB and IV). The last therapeutic strategy is radiotherapy, which is often associated with the chemotherapy [21,22].

Our study concerns patients who were at least 18 years old, residing in the region Burgundy and hospitalised between 2006 and 2008, whatever the location in France of the establishment attended. All of the patients had undergone major lung surgery for the management of non-metastatic malignant tumours.

2.2. Pathway related definitions

Given a patient, the pathway (considered the same way as care management history) was defined as the chronological succession of different discharge abstracts identified in the national PMSI database and linked together using an anonymised unique patient identifier. This pathway was characterised by four elements: first of all, the start, that is to say the first stay with surgery related to lung cancer treatment. This is often the first step in the treatment strategy. This initial management directly affects the prognosis and survival [24]. The end of the pathway occurs after one year of follow-up. We then have the list of all discharge abstracts located between the start and the end of the pathway. These discharge abstracts contain all the information recorded, notably the modifiable characteristics of patients such as the place of residence, the anti-cancer treatments given and the establishments attended. Finally, the patient's condition at the end is recorded, that is, alive or dead.

The patient pathway representation is shown as a repeat-free chronological succession of the different categories of establishments attended by a patient. The following establishment categories were used in this study: private clinics (CL), private non-profit-making clinics that are part of the public hospital sector (PSPH), hospitals (CH), teaching hospitals/or regional hospitals (CHU/CHR) and anti-cancer centres (CLCC).

For example, let us consider the following chronologically sorted list of hospital categories attended by a given patient: "CL–CL–CH–CHU–CH–CH". The corresponding pathway was represented as "CL–CH–CHU–CH", where one could notice that consecutive "CL" at the beginning or "CH" at the end was not repeated.

2.3. From database record to patient pathway: the building process

The study of the complete patient's pathway required a longitudinal approach of the chronological sequence of events. In our case, an event is likened to a single record of the administrative database. That is, a discharge abstract of a hospital stay. The first step of our work thus consisted of reconstituting the pathway for every selected patient by linking together all discharge abstracts belonging to the same patient using an anonymised patient identifier saved in each abstract. The next step was to group all the patients that shared the same individual pathway into one pathway. In this study, we are not interested in comparing the characteristics of patients but in the comparison of different pathways. Thus, the individuals of the data analysis are not the patients but the pathways. Symbolic data analysis (SDA) [19,25,26] allows us to analyse the pathways taking into account the variations of the characteristics of patients within each of them. To do this, the analysis suggests describing each pathway by the variables defined on the patients. As each pathway includes several patients, a pathway will be described, for each

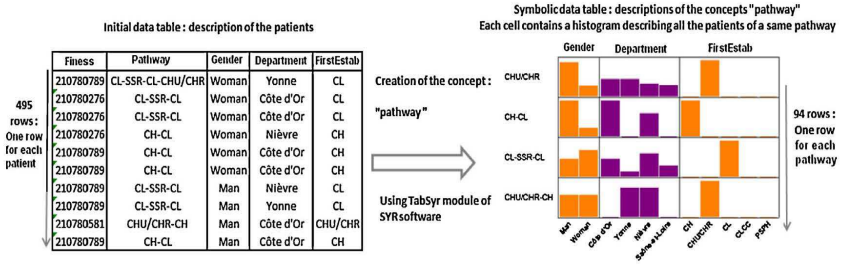


Fig. 1. From the patient's data table to the pathways data table using symbolic data analysis methodology and the SYR software.

variable, by a categorical histogram-value (histogram value) aggregating up all the different values of all the patients sharing the same pathway. Using SDA terminology, we say that we create the concept "pathway" and that we build a symbolic data table describing the pathways. Suppose a particular pathway comprises four individuals (I_1,. . .,I_4) with measurements for age (A) and gender (G). Suppose the measurements were I_1: A = 89, G = man, I_2: A = 79, G = man, I_3: A = 90, G = man and I_4: A = 76, G = woman. Then aggregating these individual patients measurements gives the pathway measurement as A = [76, 90], G = {woman (1/4), man (3/4)}. Here the age variable for pathway is now recorded as an interval value (with particular values falling across the interval), and the gender variable is recorded as a list with relative frequency (here, e.g., 3/4 for man) recorded for each of possible values (categories) in that list. Fig. 1 illustrates the construction of the symbolic data table of the pathways generated by the TabSyr module of the SYR software [27]. In this Fig. 1, 94 pathways are built from 495 patients. For instance, for a given pathway, we can consider a group of 4 patients in this same pathway living in the respective departments: Côte-d'Or, Côte-d'Or, Yonne and Nièvre. The variable department will take the following histogram-values: Côte-d'Or (1/2), Yonne (1/4), Nièvre (1/4) in which the proportions are shown in brackets.

2.4. Patient's pathway description variables

To characterise a pathway, variables were selected according to their clinical pertinence and ease of interpretation. All variables used to describe a patient's pathway were defined at the patient level. As shown in Table 1, these variables were divided in 2 groups: Group 1 contained variables that do not change across different stays. The Group 2 was concerned with global variables related to at least 2 stays. In Group 1, variables were again divided into three subgroups: the first subgroup (SG 1) included patient identification variables such as age, gender and the living department (Côte-d'or, Nièvre, Saône-et-loire or Yonne) which were recorded in the first stay of the patient's pathway and the status (alive or dead) known after the last stay. The next subgroup was formed with variables related to the first surgical operation: the year (between year 2006 and 2008 inclusive), the name of the procedure performed (lobectomy, pneumectomy, others) and the category of the establishment attended. Finally, the last subgroup was built with binary variables (yes/no) related to treatment modalities against cancer and recorded in a specific stay: the use of neoadjuvant chemotherapy, the use of chemotherapy less than 3 months after the initial surgery and completion of chemotherapy at least 3 months later. In Group 2, variables were divided into two subgroups: The first subgroup variables were used to summarised all the stays in a patient's pathway: the length of each stay (number of days), the total length of all stays, the repeat-free sequence of different categories of establishment attended and the type variable that characterised a patient's pathway as territorial (when for a given pathway, all hospitals and the patient residence were located in the same department), regional or extra-regional. The second subgroup included variables related to different treatment modalities against cancer (surgery, chemotherapy and radiotherapy): the number of received modalities,

Table 1
List of patient-related variables collected and used for the description of pathway.

Group (G)	Subgroup (SG)	Variable	Description
G 1	SG 1	Gender	1 = man; 2 = woman
		Age	The age of the patient
		Department	Patient's department of residence at the start of the trajectory
	SG 2	Status	Condition of patient at discharge from the last hospitalisation of the trajectory: alive or dead
		Year	L'année pendant laquelle a eu lieu the première hospitalisation for the première surgery for lutter contre the lung cancer
		Surgery	Represents the different thoracic surgery techniques: lobectomy, pneumonectomy, segmentectomy, partial exeresis, etc.
		FirstEstab	Category of the first establishment of the trajectory
	SG 3	Neoadjv chemo	Was chemotherapy given before initial surgery? (yes/no)
		Chemo_min3	The patient received chemotherapy less than 3 months after the initial surgery (yes/no)
		Chemo_max3	The patient received chemotherapy at least 3 months after the initial surgery (yes/no)
G 2	SG 1	Pathway	Chronological sequence without repeats for the category of establishments attended from the stay for the first surgery and throughout the following year
		Establishment	The categories of the establishments attended after the first stay for surgery
		Pathway type	Type of trajectory in 3 modalities: local, regional or regional migration
		Stay duration	Duration of stay for each hospitalisation in the trajectory
		Nbepisode	The total number of hospitalisations during the trajectory.
		Hosp time	Total number of days spent in the different hospital structures
	SG 2	Therapmod	The anti-cancer therapies used after the initial surgery
		NbTherapMod	Total number of therapies received during the trajectory. We counted one therapy per stay
		Therapy	Chronological sequence without repeats of the different therapies received during the trajectory

the complete chronological sequence of received treatment modalities (e.g., surgery–chemotherapy–chemotherapy–chemotherapy) and the repeat-free one (e.g., surgery–chemotherapy).

2.5. Pathway typology building process

In the next step, a clustering of the pathways allows us to build the main types of pathways (pathway typology) and focus on the interpretation of these main types. We applied directly to the symbolic data table of the pathways a clustering procedure that was implemented in the "CluStsyr" module of SYR software [27]. This procedure extends "*k*-means" clustering to symbolic data as input [28]. This is an iterative method that improves the homogeneity of the clusters by calculating their barycentres and by re-allocating individuals according to the new barycentres. For this type of classification, the number of clusters has to be fixed beforehand. However, it is also preferable to limit the number of variables (feature selection) used in a given clustering application both to avoid redundancy and to facilitate interpretation of results. Thus, one could consider according to their own clinical experience or experts recommendations, only, either clinically pertinence variables or those with ease of interpretation. However, we suggest a more reproducible method to perform this feature selection. That is by conducting a dimension reduction technique using a principal component analysis (PCA) technique [29,30]. The feature selection is done by selecting only one or two variables with the higher loading on one principal component (i.e., their correlation with this component or axis). The stability of the selection could be investigated using bootstrapped methods [31]. Only principal components with an eigenvalue value greater than 1.0 were retained for this variable selection process. The pathway clustering is thus performed on the selected variables called "explanatory variables". The other variables were preserved as illustrative variables for the interpretation. As output, we obtain clusters or the typology of pathways described with histogram-valued symbolic data.

For the geographical analysis, the place of residence of the patients and the location of the establishments they attended were geocoded. The different pathways were represented by spider maps made up of line segments between the patient's place of residence and the establishment identified for each episode of care management.

2.6. Data analysis results

The description of the data concerned patients' characteristics, the different treatment modalities and the different categories of establishment. Qualitative variables were described as percentages and comparisons were made using Pearson's Chi-square or Fisher's test. Continuous quantitative variables were described as means and standard deviations. Analysis of variance was used for multiple means comparison. Unless indicated otherwise, the different statistical tests were conducted with a level of significance of 5%.

The movements of patients within a given pathway were analysed in two complementary ways: the first was when, by using the notion of transition between two types of establishment, we calculated the different transition rates. In the second way, an analysis was done in terms of movement of the patient around the country, from the department of residence to an establishment. From here we were able to classify each pathway in one of the following 3 categories: first, the so-called territorial pathway, in which patients only attended establishments located in their department of residence (proximity pathways). Second, was the regional pathway where the hospital attended was located outside the department of residence, but still within the region of Burgundy. Finally, we had a regional migration pathway where at least one hospital stay occurred outside the region of Burgundy.

Data were extracted from the national PMSI databases using a dedicated extraction tool, which had been designed and developed for this study. The reconstitution of pathways as well as their description was done using SAS version 9.2 software. The tables for symbolic data and the classification of pathways were constructed using the ClustSYR software suite from SYROKKO Company. The MapInfo 8.5 and Adobe Illustrator CS 3 software was used to create the cartographic representations of the different pathways types constructed.

3. Results

First we will present the characteristics of the patients and those of the reconstituted hospital pathways before and after removal of repeats. We will then present the results of the classification of these pathways as well as the spatial illustration.

We identified 495 patients who met the defined inclusion criteria between 2006 and 2008. These patients accounted for a total of 3831 hospital stays with all types of care included in a time window of one year following the first tumour resection. The majority of patients were men 79% (393) versus 21% (102) for women. There was no significant relationship between gender and department of residence at the first hospitalisation ($p = 0.253$). For age, we found no significant difference between men (64; 95% confidence interval (CI) 63–65) and women (63; 95% CI 61–65), at the start of the pathway and according to the department of residence ($p = 0.404$). The mean number of hospital stays per patient was not significantly linked to either gender, or place of residence at the start of the trajectory ($p = 0.640$). It was 8 (95% CI 7–9) hospital stays for men and 8 (95% CI 7–8) hospital stays for women (cf. Table 2).

Individual pathways for the 495 patients were reconstituted. These comprised on average 9 episodes of care with a median of 6 and a standard deviation of 7. The longest pathway comprised 78 episodes (hospital stays) and the shortest 1. Care was given in 6 different types of establishment: Private Clinics (CL) had the highest attendance rates at 41.6% followed by Hospitals (CH) with 31.2%, then Teaching Hospitals (CHU) with 16.3%. Under the 10% threshold, there were PSPH at 6%, Anticancer Centres (CLCC) at

Table 2
Number of patients and certain individual characteristics by department of residence.

Variables	Côte d'Or	Nièvre	Saône-et-Loire	Yonne	Total	P[a]
No. of patients	111	85	187	112	495	
Gender						
Men	85	63	150	95	393	0.253
Women	26	22	37	17	102	
Age[b]	63 (61–65)	63 (61–65)	64 (63–66)	64 (63–66)	64 (63–65)	0.404
No. of episodes[b,c]	8 (7–9)	6 (5–8)	8 (7–9)	9 (6–10)	8 (7–8)	0.640

[a] Significance.
[b] Mean (95% confidence interval).
[c] Number of stays in an establishment.

Table 3
The 10 most frequent raw pathways.

Pathway	Frequency	Proportion (%)
CHU[a]/1	37	7.5
CHU/1-CH/1[b]	16	3.2
CL/2	13	2.6
CL/3	13	2.6
CHU/1-CL/1	12	2.4
CL/1	12	2.4
PSPH/1	11	2.2
CHU/1-CH/2	10	2
CHU/2	8	1.6
CHU/1-CH/12	7	1.4

[a] CH, hospital; CHU, teaching hospitals; CL, private clinics; CLCC, anticancer centres; PSPH, private non-profit-making clinics, part of the public hospital sector.
[b] CH/*y*, *y* = number of repetitions of the type of establishment in the trajectory.

Table 4
Distribution by department of the different types of establishment in the region of Burgundy.

Type of establishment	Côte-d'Or	Nièvre	Saône-et-Loire	Yonne	Total
CHU/CHR	1				1
CH	1	2	5	2	10
CLCC	1				1
CL	5	1	4	2	12
Total	8	3	9	4	24

4.7% and establishments for follow-up care (SSR) at 0.2%. The 10 most frequent pathways accounted for about 30% of the total. Table 3 shows these different pathways represented here in the repeat-free form, in which the number of repeats is shown after the slash. The most frequent pathway occurred in 7.5% [37] of cases with a single type of establishment (CHU repeated once). Then came the association CHU–CH which accounted for 3.2% [16], etc. Altogether, a total of 54 establishments were attended including 44% [24] located in the region of Burgundy. Table 4 shows the distribution of these 24 types of hospitals in the four departments of this region. It is noted that neither PSPH, nor SSR was represented in Table 4. This is because follow-up did not take place in these types of establishments located in region Burgundy even though they exist.

From these pathways, we were able to analyse the movements of patients. We thus calculated the different transition rates, which are shown in Table 5. We found that the vast majority of transitions occurred between identical types of establishment (CHU → CHU, for example). In addition, the highest rates were towards Hospitals (CHU → CH (24%) and PSPH → CH (22%)). With regard to the patient's movements around the country, the proximity pathways accounted for 41% (203) of patients. The regional one for 15% (74) of patients and the regional migration pathway was the most frequent with 218 patients (44%).

The symbolic data table contained 94 distinct pathways. The mean number of patients per pathway was 5 (95% CI 3–8) varying from 1 to 83. These pathways were described by the 18 variables, summarised in Table 1. These variables are all presented in the form of categorical histograms with the number of modalities per variable which range from 2 (gender) to 6 (age group). When the different modalities of these variables are added together, we obtain approximately 54 types of information that correlate more or less with each other.

With the PCA, 22 principal components that accounted for 84% of the information were identified. For each component (axis), the higher loading variables (i.e., their correlation with this axis) as shown in Table 6 and highlighted in bold, were candidates for our explanatory set of variables. Thus, we see our 18 variables were principally correlated with 10 of these principal components. On the first axis (axis 1), 5 variables were related to treatment

Table 5
Transfer rates between the types of establishment.

From/To	[→ CH]	[→ CHU]	[→ CL]	[→ CLCC]	[→ PSPH]	[→ SSR]
[CH →]	0.89	0.03	0.05	0.02	0.02	0.00
[CHU →]	0.24	0.57	0.14	0.03	0.03	0
[CL →]	0.04	0.03	0.92	0.01	0.01	0.00
[CLCC →]	0.10	0.05	0.06	0.80	0	0
[PSPH →]	0.22	0.04	0.16	0.02	0.56	0
[SSR →]	0	0	0.13	0	0	0.87

Table 6
Correlation coefficients between the variables that describe pathways and the principal components from the principal component analysis with eigenvalue >1.

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Axis 8	Axis 9	Axis 10	Axis 11
Gender	0.077	0.104	0.272	0.344	0.268	**0.439**	0.196	0.371	0.085	0.005
Chemo min3	**0.868**	0.025	0.016	0.078	0.087	0.025	0.014	0.002	0.064	0.026
Chemo max3	**0.831**	0.123	0.074	0.087	0.146	0.097	0.053	0.082	0.047	0.114
Neoadjv chemo	0.01	**0.587**	0.227	0.03	0.075	0.257	0.057	0.092	0.454	0.147
Pathway type	0.346	0.565	0.444	0.399	**0.621**	0.091	0.109	0.026	0.163	0.079
Tps hosp	0.322	0.628	**0.714**	0.444	0.502	0.277	0.203	0.377	0.361	0.152
FirstEstab	0.346	0.489	0.455	**0.674**	0.642	0.625	0.606	0.205	0.254	0.139
Department	0.358	0.586	0.419	**0.734**	0.401	0.472	0.359	0.163	0.454	0.299
Year	0.139	0.226	0.072	0.172	0.276	0.349	0.208	0.291	0.215	**0.541**
Age	0.433	0.394	0.23	0.396	0.226	0.251	0.52	0.531	**0.661**	0.476
Surgery	0.107	0.347	0.487	0.28	0.201	**0.528**	0.195	0.299	0.127	0.157
Status	0.089	0.311	**0.444**	0.024	0.317	0.328	0.406	0.187	0.005	0.351
NbTherapMod	**0.949**	0.452	0.239	0.392	0.257	0.379	0.182	0.442	0.151	0.092
Nbepisode	0.544	0.433	0.477	0.275	0.265	0.361	0.365	**0.609**	0.346	0.235
Therapy	**0.928**	0.282	0.124	0.035	0.138	0.151	0.104	0.13	0.086	0.035
Stay duration	0.439	0.529	**0.756**	0.519	0.504	0.24	0.346	0.459	0.344	0.285
Establishment	0.404	0.396	0.437	0.483	0.367	0.284	**0.558**	0.2	0.479	0.495
Therapmod	**0.84**	0.12	0.122	0.103	0.085	0.138	0.244	0.135	0.094	0.031

Variables in bold were selected for the classification.

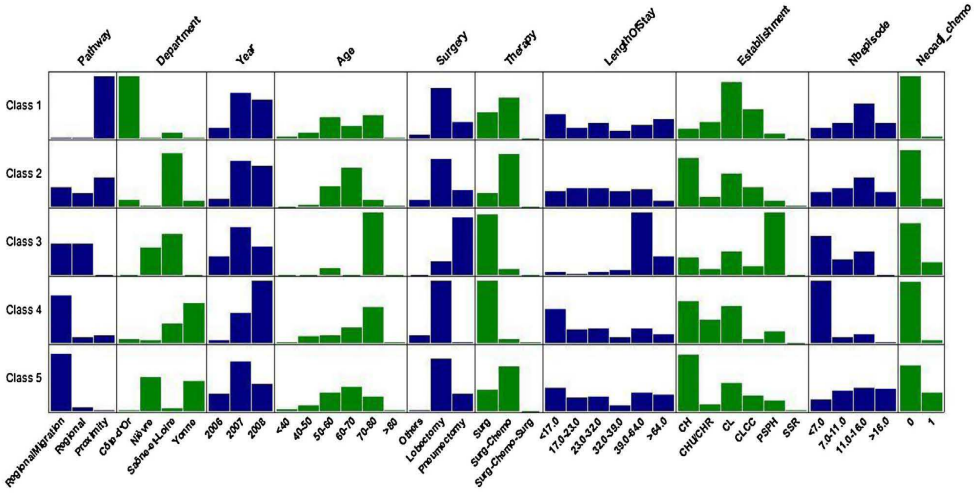


Fig. 2. Description of the 5 classes (profiles) using pathway description variables.

procedure. Axis 3 extracts information on the duration of hospital stays with 3 variables, and provides information on the status of the patient at the end of the trajectory (alive or dead). Axis 6 gives information principally on the gender of patients and on the type of surgery (lobectomy, pneumonectomy or other). In contrast, other axes are more specific to one variable, such as axis 5 for the type of trajectory (proximity, regional or regional migration), axis 10 for age and axis 11 for the year the disease was diagnosed. By balancing these results across all axes and all variables, a single variable per axis was selected to form our explanatory variable set used for the clustering process. The retained set of variables represented the most frequent ones selected using the bootstrap method.

Fig. 2 shows the results of the typology with 5 distinct types of pathway. In this figure, clusters are listed in lines while each column represents variables describing pathways. The names of the variables are shown in the upper part of the figure and the different modalities in the lower part. The first six variables represented in Fig. 2 were part of our "explanatory variables" list and were less discriminant from left to right. The last three variables were other pathways descriptive variables. Every cell in the figure shows a categorical histogram of the different modalities of a variable in a given class. The higher the bar, the more this modality is represented in the corresponding class.

The first column suggests that there are three main types of pathway: migration outside the region (regional migration of patients) (classes 4 and 5), care exclusively within the department of residence (proximity: classes 1 and 2) and care outside the patient's department of residence (class 3). This last class includes as many migrations as intra-regional care; 60% of the patients concerned lived in Saône-et-Loire and 40% lived in the Nievre. Most were elderly men (70–80), who had surgery alone and, unlike patients in other groups, underwent pneumonectomy. This operation is performed essentially in Teaching Hospitals (First-Estab), and follow-up principally takes place in PSPH. Regional migrations alone principally concern patients from Yonne (class 4) and Nievre (class 5). For the former, the treatment was surgery alone (lobectomy) performed in a PSPH establishment and for the latter, the treatment was a mixture of surgery and chemotherapy. It is noteworthy that for this cancer, the gender of the patient did not make it possible to distinguish between the different classes or the state of the patient at the end of the pathway (living patients predominated).

The cartographic representation of the classification made it possible to refine the interpretation of the first results (cf. Fig. 3). One outstanding result is the extent of migrations outside the region of Burgundy, for class 5 patients, who travelled to Paris and Lyon.

More generally, the cartography of the pathway types shows the strong polarisation towards Dijon CHU, which draws patients from the whole region. This representation brings to mind models concerning the attraction of urban centres and their rural outskirts, reflecting not only population dynamics and economic power, but also utilisation of the healthcare system. Patients who live in rural environments are those who cover the longest distances to reach large urban centres for their surgical operations [32–34].

To refine the interpretation of the classification table, we created Table 7, which summarises the number of patients, the number of pathways and the ratio between the two, according to the different classes. It is noticeable that the number of patients in class 3 appears to be smaller than those in the other classes.

4. Discussion

We reconstructed and described hospital pathway for each patient. We then identified five distinct patient profiles (pathways), which allowed us to make a synthesis of the longitudinal evolution of Burgundy patients operated on for lung cancer and followed for 1 year, using a process of classification based on variables available in the PMSI database: the type of establishment for the first surgery, the type of pathway (proximity regional or migration outside the region – called regional migration), age, the department of residence, the nature of the surgery (lobectomy, pneumonectomy, other), the establishment attended, the number and duration of hospital stays. The different profiles are easy to understand. The first profile (class 1) includes patients who make use of local care providers (proximity). They have their operations and chemotherapy treatment in private clinics. These are principally men living in the department of Côte-d'Or and aged between 50 and 80 years (with relatively well-balanced 10-year age groups). Class 2 mainly includes men between 60 and 70 years old living in Saône-et-Loire. Most are treated in the department, but some travel to other departments of the region or even beyond. There is the same variability with regard to the establishment where the surgery is performed, with a preference for CHU/CHR, whereas chemotherapy is administered in CH. The last three

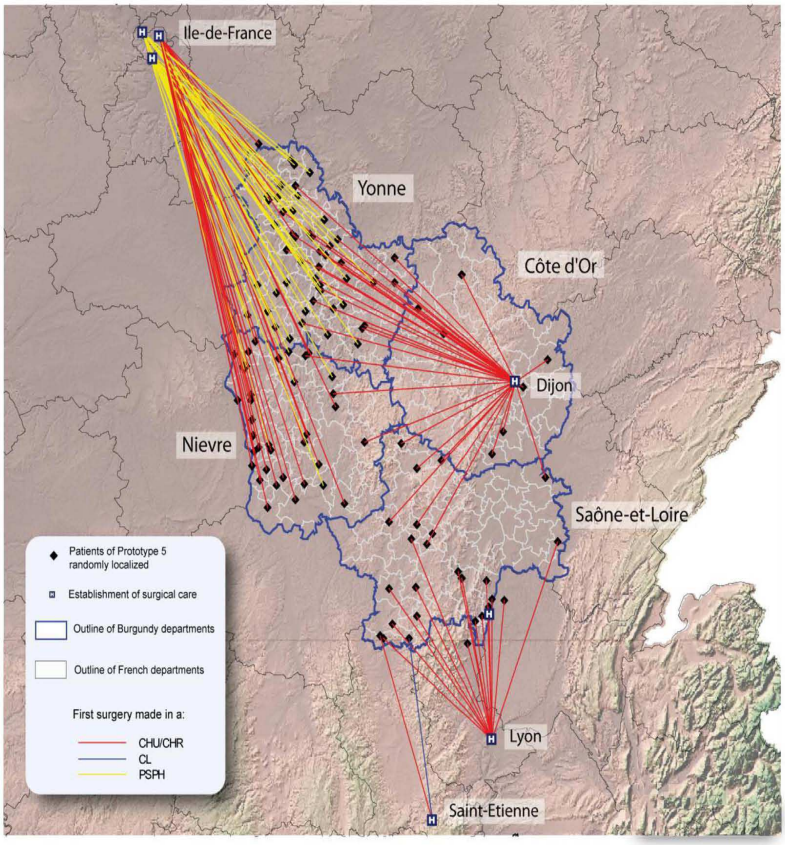


Fig. 3. Cartographic representation of class 5.

Table 7
Description of the 5 classes (profiles) resulting from the classification in terms of number of patients, pathways and the patients/pathways ratio.

	Number of patients	%	Number of pathways	%	Ratio patient/pathways
Class 1	141	28.5	15	16.0	9
Class 2	77	15.6	30	31.9	3
Class 3	6	1.2	5	5.3	1
Class 4	131	26.5	15	16.0	9
Class 5	140	28.3	29	30.9	5
Total	495	100.0	94	100.0	5

profiles (classes 3–5) essentially concern patients who are treated outside the region of residence. Class 3 includes mainly elderly men (70–80 years) who have their operations in CHU/CHR for pneumonectomy and are then followed in a PSPH. Most live in Nievre and a few in Saône-et-Loire. Most of the patients in classes 4 and 5 are managed almost exclusively outside this region of residence. Class 4 includes elderly patients (70–80 years) from Yonne, but also from Saône-et-Loire who have their operations in PSPH and who almost never have chemotherapy after the lobectomy. This class is the one that contains the largest proportion of women. Patients in class 5 are slightly younger, have their operations in CHU/CHR and PSPH and their chemotherapy in CH. Spatial analysis of these profiles provided a clear picture of the destinations of patients who seek treatment outside their region. These account for a large proportion of all pathways (44% of pathways correspond to regional migrations). The majority of patients in class 4, for example, were operated in PSPH in the region Île-de-France. These results are important for healthcare policy makers in their struggle to reduce inequality in access to healthcare services and improve equity in service provision. In fact, this classification has highlighted 2 groups (4 and 5) of patients whose pathways are the most likely to include migration outside residential region for a specific healthcare service. Hence, identification of associated factors is more precise.

The methodology used in this study has several strong points: the choice of the type of epidemiological study (ecological), the analytical tools employed, which are innovative in the field of epidemiology [25,26,35,36] and finally the very nature of the data used for the application. These medico-economic databases are particularly interesting for the wealth of individual information they contain, their volume and diversity [7,8]. The efficacy of certain healthcare policies can thus be evaluated thanks to ecological studies. In our case, the choice of the type of study was in no way a constraint. To the contrary, our research was driven by the desire to extract as much as possible of the wealth of individual data available. This is as important for policy

decision-makers as for clinicians because this strategy makes it easier to understand and visualise the groups analysed. The essential aim here was to minimise information loss during the aggregation process [25]; that is to say the transfer of individual data (patients) to pathway data. The work that we have done on the management of lung cancer in the region of Burgundy consisted of using, as the principal source of data, information on patients' individual hospital stays and providing as the final result a photograph of types of care management (profiles).

The cartography of the results provides a dynamic spatial view of patients' pathways in each of the classes. Although not all of the patients' pathways can be represented here, by using spider maps to show the pathways, it is easier to understand the destination of each of the patients according to the department of residence. The model used is not limited to surgical management alone, but can also be generalised for all aspects of care. The aim is to appreciate the spatial footprint of these pathways. It must be pointed out that this method of spatial analysis is based on a limited number of variables: identification number of the establishments and postcode of the patients' residence. Another advantage of this methodology lies in the interpretation of the classification results, which is based on classical individual variables (age, gender, establishments attended, etc.), which are then aggregated. In addition, the representation in the form of categorical histograms makes it possible to see within pathways variability for each variable. Because they are simple and easy to interpret, these results are a useful tool to help decision-making for planners and healthcare professionals.

Of course, although analytic methods can generate simple, easy-to-interpret results, none is perfect, and our work does not escape this rule. Debatable choices were made throughout the study. The definition of the notion of pathway in our context was important. We chose initial surgery as the start of the pathway for all of the patients because the information was available for all, relatively easy to collect and presented no ambiguity in the coding. The duration of the pathways, which was set at 1 year after the initial surgery, was adapted to the periods covered by the databases we used, as was the case for all of the automatic classification issues, which rely on partitioning methods [20], and the choice of the number of classes was a compromise. In the spatial analysis phase, the identification and location codes were easy to geocode, but only gave an approximate idea of the location of the patients and the structures. We can give the example of 'Assistance Publique – Hôpitaux de Paris' establishments, which were geocoded for the address of the organisation's head office; in the same way, patients were located randomly in the PMSI zones. Nonetheless, the representation of the surgical management pathways by line segments symbolising the distances travelled, did not have an impact on the general perception of migrations nor on the dynamics of hospitals' ability to attract, even though the use of this technique made it impossible to include in the analysis a variable linked to the means of transport used.

In the scientific medical literature, the desired information is at the level of the patient. Given the volume of national medico-administrative databases, it can be advantageous to aggregate data to work with more compact datasets. This study can be considered an ecological study without the drawbacks (Ecological fallacy) [37], because it is always possible to come back to a detailed level. A similar question arises with regard to the management of a huge number of variables (sometimes several hundred). Should the number of variables be reduced, with the risk of leaving out important variables? The interest of factorial analysis lies in the fact that all of the variables in the model are taken into account, while transforming the information into a number of components that is smaller than the initial number of variables.

5. Conclusion

The combination of the methods employed in this study made it possible to synthesise in a simple and reproducible manner the mass of information contained in medico-administrative databases. Actually, the data provide concrete knowledge about the organisation of care management in the field in this context of a serious disease like lung cancer. By using appropriate tools, the construction of pathways in care management allows better understanding of the logic that governs patients' movements, as well as characteristics of the therapeutic episodes they have experienced. Though implementing the tools may sometimes appear complex, interpretation of the results remains simple, and the main profiles of patients and their pathways revealed by the classification make it possible to extract the sparse content of hospital databases. In addition, the combination of statistical analyses and cartography provides an overall trans-disciplinary view of public health. The reconstruction of care-management pathways opens the way for analyses of collaboration between the different healthcare establishments, which could reveal associations that should be developed. We can suppose that with the generalisation of the use of medico-administrative databases, in France as well as in Europe, the employment of such methods in the field of public health will develop and will provide numerous elements useful to decision-makers and to the implementation of efficient healthcare governance.

Conflicts of interest statement

All the authors declare they have no financial or personal relationships with other people or organizations that could inappropriately influence or bias their work.

References

[1] Department of Health. A national cancer strategy for the future. Stockholm: Department of Health, 2009.
[2] Department of Health. Improving outcomes, a strategy for cancer. United Kingdom: Department of Health, 2011.
[3] Steger C, Daniel K, Gurian GL, Petherick JT, Stockmyer C, David AM, et al. Public policy action and CCC implementation: benefits and hurdles. Cancer Causes Control 2010;21(12):2041–8.
[4] Institut National du Cancer. Plan cancer 2009–2013. France: Institut National du Cancer, 2009. 140.
[5] Walter S. Economic control of quality of manufactured product. ASfQCD; 1980. 501.
[6] Deming WE. In: Press M, ed. Out of the crisis. MIT Press; 2000: 523.
[7] Roos NP, Roos LL, Brownell M, Fuller EL. Enhancing policymakers' understanding of disparities: relevant data from an information-rich environment. Milbank Q 2010;88(3):382–403.
[8] Roos LL, Menec V, Currie RJ. Policy analysis in an information-rich environment. Soc Sci Med 2004;58:2231–41. England.
[9] Roger France FH. Case mix use in 25 countries: a migration success but international comparisons failure. Int J Med Inform 2003;70(2–3):215–9.
[10] Roos Jr LL, Nicol JP, Cageorge SM. Using administrative data for longitudinal research: comparisons with primary data collection. J Chronic Dis 1987;40(1):41–9.
[11] Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health 2011;32:91–108.
[12] Akushevich I, Kravchenko J, Akushevich L, Ukraintseva S, Arbeev K, Yashin AI. Medical cost trajectories and onsets of cancer and noncancer diseases in US elderly population. Comput Math Methods Med 2011;857892.
[13] Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. Methods Inf Med 1998;37(3):271–7.
[14] Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J, Dusserre L. How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure. Int J Med Inform 1998;49(1):117–22.
[15] Quantin C, Fassa M, Coatrieux G, Trouessin G, Allaert FA. Combining hashing and enciphering algorithms for epidemiological analysis of gathered data. Methods Inf Med 2008;47(5):454–8.
[16] Megan AB, Damien J, Vijaya S, Sue E, David VP, Ian S, et al. Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 2010;10(346).

[17] Olive F, Gomez F, Schott AM, Remontet L, Bossard N, Mitton N, et al. Critical analysis of French DRG based information system (PMSI) databases for the epidemiology of cancer: a longitudinal approach becomes possible. Rev Epidemiol Sante Publique 2011;59(1):53–8.
[18] 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(1):27–34.
[19] Billard L, Diday E. Symbolic data analysis: conceptual statistics and data mining. Wiley; 2006.
[20] Edwin D. Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. Rev Stat Appl 1971;19(2):19–33.
[21] Bizieux-Thaminy A, Hureaux J, Hureaux T. Cancers bronchiques primitifs: bilan diagnostique et traitement (primary lung cancer: diagnostic and treatment). EMC Med 2004;1(1):8–17.
[22] De Leyn P, Decker G. Le traitement chirurgical du cancer bronchique non à petites cellules. Rev Maladies Respir 2004;21(5):971–82.
[23] INCa. Recommandations professionnelles Cancer du poumon non à petites cellules: prise en charge thérapeutique. Boulogne-Billancourt: INCa, 2010.
[24] Jacques D, François G, Gérard D, Laurent B. Le livre blanc de la chirurgie cancérologique. Bull Cancer 2002;89(10):2.
[25] Diday E. Introduction a l'analyse des donnees symboliques; 1989;38.
[26] Diday E, Noirhomme-Fraiture M. Symbolic data analysis and the SODAS software. Wiley; 2008. 476.
[27] Afonso F. User manual of the SYR software. Publication Si; 2012.
[28] Carvalho FAT, Lechevallier Y, Verde R. Clustering methods in symbolic data analysis. In: Diday E, Noirhomme-Fraiture M, eds. Symbolic data analysis and the SODAS software. John Wiley & Sons Ltd; 2007: 181–203.
[29] Makosso-Kallyth S, Diday E. Adaptation of interval PCA to symbolic histogram variables. Adv Data Anal Classif 2012;6(2):147–59.
[30] Diday E. Principal component analysis for bar charts and metabins tables. Stat Anal Data Mining 2013;11188. http://dx.doi.org/10.1002/sam.
[31] Efron B. The jackknife the bootstrap and other resampling plans. SIAM Publishers; 1982.
[32] Ahamad A. Geographic access to cancer care: a disparity and a solution. Postgrad Med J 2011;87(1031):585–9.
[33] Onega T, Duell EJ, Shi X, Wang D, Demidenko E, Goodman D. Geographic access to cancer care in the US. Cancer 2008;112(4):909–18.
[34] Riva M, Curtis S, Gauvin L, Fagg J. Unravelling the extent of inequalities in health across urban and rural areas: evidence from a national sample in England – Discover – Canada Institute for Scientific and Technical Information. Soc Sci Med 2009;10.
[35] Diday E. Quelques aspects de l'analyse des données symboliques; 1993.
[36] Quantin C, Billard L, Touati M, Andreu N, Afonso F, Battaglia G, et al. Classification and regression trees on aggregate data modeling: an application in acute myocardial infarction. J Probability Stat 2011.
[37] Hart J. On ecological studies: a short communication. Dose Response 2011;9(4):497–501.

Article original / Original Article *Journal d'Économie Médicale 2012, Vol. 30, n° 2, 00-00*

Analyse spatiale des trajectoires de prise en charge des patients atteints de cancer primitif du poumon en région Bourgogne

A. ROUSSOT[1,3], E. COMBIER[3], G. NUEMI[1,2], J.M. AMAT-ROZE[4], C. QUANTIN[1,2]

[1]CHRU, Service de Biostatistique et d'Informatique Médicale, CHU de Dijon, France

[2]Inserm, U866, Dijon, F-21000, Université de Bourgogne, Dijon

[3]Centre d'épidémiologie des populations, EA 4184, Université de Bourgogne, Dijon

[4]Université Paris-est Créteil, Lab'URBA EA 34 82

RÉSUMÉ

Ce travail s'inscrit dans les champs de la recherche en santé publique et en géographie de la santé et propose une approche multiniveau d'analyse des trajectoires de prise en charge de patients bourguignons atteints de cancer du poumon. Alors que les établissements sanitaires sont soumis à des seuils d'activité pour pratiquer certaines interventions chirurgicales dans le cadre de prise en charge oncologique, il était nécessaire de mettre au point une méthode d'analyse de ces trajectoires. De plus, les flux de patients qui accompagnent ces trajectoires en cancérologie engagent souvent des déplacements qui répondent à certaines logiques spatiales. A partir de la base nationale du PMSI, on a sélectionné 416 patients atteints de cancer du poumon en région Bourgogne dont les trajectoires de soins ont été classifiées via des analyses cartographique et statistique. Les trajectoires débutent avec un acte chirurgical majeur et correspondent à une succession d'épisodes de prise en charge sur une durée de un an. La répartition spatiale de la patientèle suit les grands foyers de peuplement bourguignons et les grands axes de polarisation. Les flux de patients s'affranchissent des multiples découpages institutionnels et sanitaires et créent des liens entre les territoires de la région d'étude et les pôles sanitaires bourguignons, franciliens et lyonnais. Pour conclure, les bases PMSI sont un matériel précieux pour l'étude des trajectoires des patients et la mise en lumière des dynamiques interterritoriales engendrées par le recours aux soins. Les flux et l'orientation des trajectoires sont autant d'observations utiles pour l'élaboration d'une gouvernance sanitaire pragmatique.

Mots-clés : PMSI, analyse spatiale, trajectoires de soins, planification sanitaire, cancer.

SUMMARY

SPATIAL ANALYSIS OF TRAJECTORIES OF CARE OF PATIENTS REACHED BY PRIMITIVE LUNG-CANCER IN BURGUNDY REGION

This work introduces a multi-level analysis of health care trajectories of patients reached by lung cancer in Burgundy, France. This study combines researches in public health as well as geography of health and relies on the use of the French DRG database. The French national Cancer plan, applied since 2003, imposes a minimum threshold of surgeries on the health structures. As the Organization of health care is going to change, it seemed important to study spatial logics which model the patient's travels during their care trajectories. We used the French DRG database to select 416 patients suffering from primary lung cancer and reconstruct their health care trajectories. We defined these trajectories as the succession of every hospital stays recorded for each patient; the trajectories begin with a thoracic surgical intervention. Patients and their trajectories were classified with cartographic and statistical analysis. The maps show that patients are localized in the metropolis and along the axis of polarization. The flows of patients throughout regional and administrative boarders link up Burgundy to care structures in Paris and Lyon. Patients from the North head for Paris, whereas patients from the South of Saône-et-Loire go to Lyon. The teaching hospital of Dijon attracts patients from the entire region. To conclude, DRG databases become a strong material even more using in public health researches, the medical information they contain is targeted to economists, doctors, but also geographers. Studying patients' floodtides and trajectories supplies decision-makers with useful observations in order to build insightful sanitary governance.

Keywords: *DRG, spatial analysis, care trajectories, sanitary planning, lung cancer.*

1. INTRODUCTION

Une des hypothèses étiologiques de l'inégalité face au cancer est la géographie de l'offre de soins. Pour répondre au défi d'une prise en charge de qualité pour tous les patients quelle que soit leur porte d'entrée dans le système de soins, un dispositif d'autorisations a été mis en place. Désormais, les établissements hospitaliers doivent respecter des seuils d'activité en chimiothérapie, radiothérapie externe, et chirurgie, et notamment pour l'octroi d'une autorisation à exercer des interventions de chirurgie thoracique chez les patients atteints de cancer du poumon. Dans ce contexte, un état des lieux des trajectoires de soins des patients et de leurs typologies par type de cancer permettrait d'avoir une description de l'existant (en termes de fréquentations d'établissements par les patients) et apporterait des éléments qui pourraient être utilisés lors de la prochaine campagne d'autorisation.

Peu de travaux traitent de l'élaboration et des caractéristiques des trajectoires de soins, malgré une proposition dans le domaine des soins psychiatriques [1], un secteur de prise en charge sanitaire particulier, dont les déterminants ne peuvent s'appliquer à la cancérologie. Ce travail se place donc dans le champ de l'expérimentation, surtout pour le volet de l'analyse spatiale.

2. HYPOTHÈSES ET OBJECTIFS

On peut faire l'**hypothèse** que la pratique d'un territoire de santé engendre des logiques de déplacement singulières, parfois ancestrales, qui reposent à la fois sur la sédimentation historique, sur la densité du maillage sanitaire [2], ainsi que sur la proximité des équipements. **Les objectifs** de notre étude étaient 1) de rechercher les logiques d'utilisation de la structure hospitalière de premier recours et -2) de visualiser par des études cartographiques l'emprise spatiale des trajectoires des patients.

3. MATÉRIELS ET MÉTHODES

Population étudiée

Il s'agit d'une étude rétrospective effectuée à partir des bases nationales du PMSI (Programme de Médicalisation des Systèmes d'Information). Etaient éligibles les patients âgés d'au moins 18 ans, résidant dans la région Bourgogne, dont la prise en charge pour cancer primitif du poumon a débuté en 2006, 2007 ou 2008. Le repérage des malades a été fait à partir d'une liste de 19 actes chirurgicaux du répertoire CCAM (Classification Commune des actes Médicaux) directement liés au traitement d'un cancer primitif du poumon. La validité de la base après extraction a été étudiée en vérifiant que les diagnostics principaux des séjours concernés correspondent bien à des diagnostics de cancer primitif du poumon (codes CIM-10). La sélection des actes CCAM et des codes CIM-10 utilisés comme référence pour l'extraction et la validation a été faite à dire d'expert.

La population de l'étude est composée de 416 patients, 328 hommes et 88 femmes âgés de 38 à 85 ans, dont l'acte chirurgical inaugural de la maladie a été pratiqué dans 28 établissements répartis sur l'ensemble du territoire français.

4. MÉTHODES

Une analyse cartographique des flux des patients de leur domicile vers les lieux d'hospitalisation, ainsi qu'une représentation des classes de trajectoires ont été réalisées. Ces trajectoires concernent le premier acte de chirurgie thoracique et les recours suivants des patients. Les trajectoires correspondent à la totalité des actes vécus par un patient sur une durée de un an à partir de la première opération chirurgicale de traitement d'un cancer du poumon. Le géocodage des domiciles des patients correspond à celui des centroïdes du

zonage PMSI des lieux de résidence, celui des établissements a été réalisé à l'adresse, à partir de leur numéro FINESS juridique.

L'extraction des données des bases nationales du PMSI, la vérification des données et les analyses statistiques ont été effectuées à l'aide du logiciel SAS 9.2.

Les cartes et les diagrammes en oursin ont été réalisés avec le logiciel MapInfo® et ont été retravaillés avec le logiciel de dessin vectoriel Adobe Illustrator®. L'outil « spider graph » de MapInfo® a été utilisé pour connaître les distances euclidiennes parcourues entre la zone PMSI de résidence et l'établissement fréquenté pour un acte de chirurgie, ces distances exprimées en kilomètres ont permis de réaliser un score pour chaque patient. La distribution en quartiles de ces distances permet de synthétiser cette information et de la découper en quatre classes qui serviront de base pour l'élaboration du score. La distance parcourue par les patients est classée « faible » si celle-ci est comprise entre 4,2 et 22,3 kilomètres, « modérée » entre 22,3 et 85,9 kilomètres, « élevée » entre 85,85 et 141,9 kilomètres et « très élevée » lorsqu'elle est supérieure à 141,9 kilomètres. Pour faciliter l'analyse des correspondances, le nombre de modalités a été réduit au maximum, ainsi, la localisation des structures fréquentées a été simplifiée et reprend seulement le département: Yonne, Saône-et-Loire, Nièvre, Côte-d'Or, Lyon, Ile de France (IDF) et Autre. Le territoire de résidence de chaque patient est résumé par le numéro du département. La Saône-et-Loire a été coupée en deux parties nord et sud, afin de juger de la spécificité des trajectoires des patients de ce département.

Les trajectoires de prise en charge des cancers dans la région Bourgogne ont été classées selon une extension de la méthode des nuées dynamiques par les données symboliques [3, 4] grâce au logiciel SYR.

5. RÉSULTATS

Description générale des trajectoires

Les 416 patients sont répartis sur l'ensemble de la Bourgogne mais les effectifs par département varient d'un département à l'autre (Côte d'Or : 74 ; Nièvre : 78 ; Saône et Loire : 165 ; Yonne : 89). La répartition spatiale de la patientèle suit les grands foyers de peuplement bourguignons et les grands axes de polarisation de la région: vallées de l'Yonne, de la Loire et de la Saône, bassin minier et métropoles départementales. On observe de faibles effectifs de patients dans le Morvan et sur le Plateau bourguignon ; on peut même relever que certains territoires sont épargnés, comme le Tonnerrois à l'est de l'Yonne et le Charolais au sud-ouest de la Saône-et-Loire. A l'inverse, trois pôles présentent un effectif de patients relativement important : Sens, Dijon et Chalon-sur-Saône (cf. Figure 1).

Pour le premier acte de chirurgie, le choix des patients suit une logique de déplacement vers les grands établissements régionaux. Trois formes de polarité apparaissent :

- Polarité de certains centres urbains départementaux qui accueillent des centres hospitaliers ou des cliniques.
- Polarité régionale et rayonnement de Dijon.
- Polarité exogène de Paris et de Lyon.

Les flux de patients suivent donc une hiérarchisation du réseau. On note toutefois un défaut d'attraction de certains centres urbains, comme Mâcon qui se trouve dans l'aire d'influence lyonnaise, ou Auxerre. Les fuites hors Bourgogne vers l'est et l'ouest sont mineures. La fréquentation des établissements franciliens est surtout le fait des patients icaunais et nivernais.

On remarque l'absence de tropisme des Nivernais pour le pôle dijonnais (cf. Figure 2).

Figure 1

Localisation des patients atteints de cancer du poumon (2006-2008)

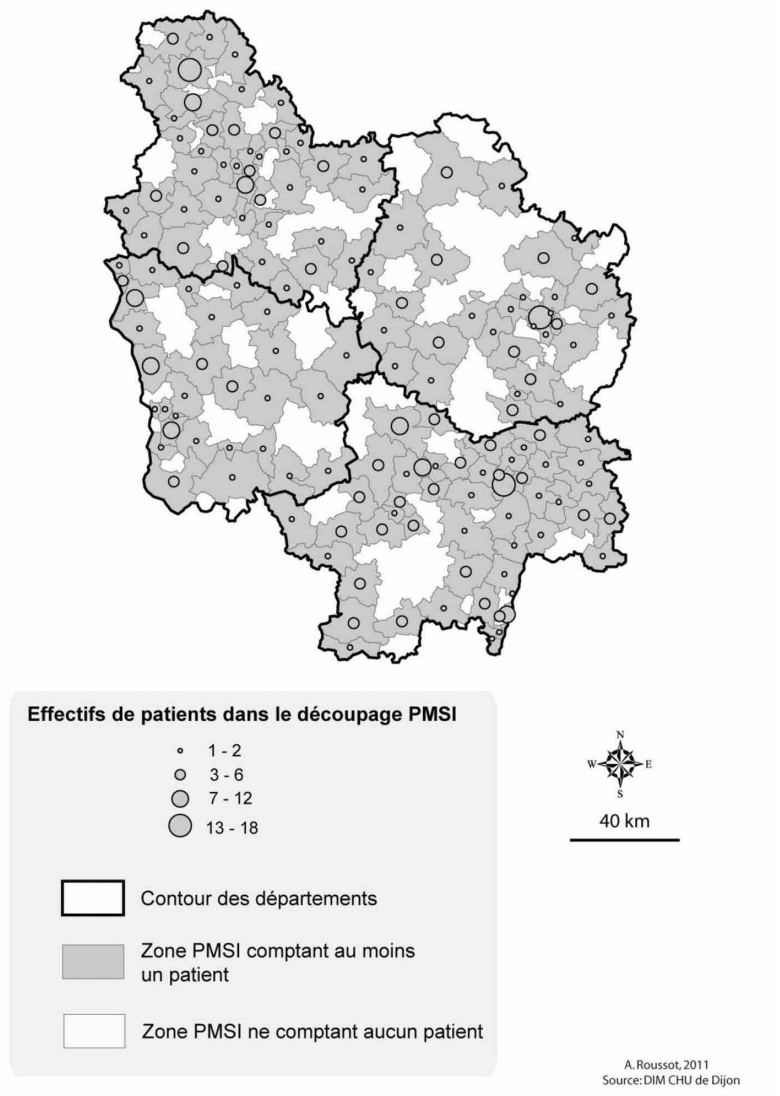
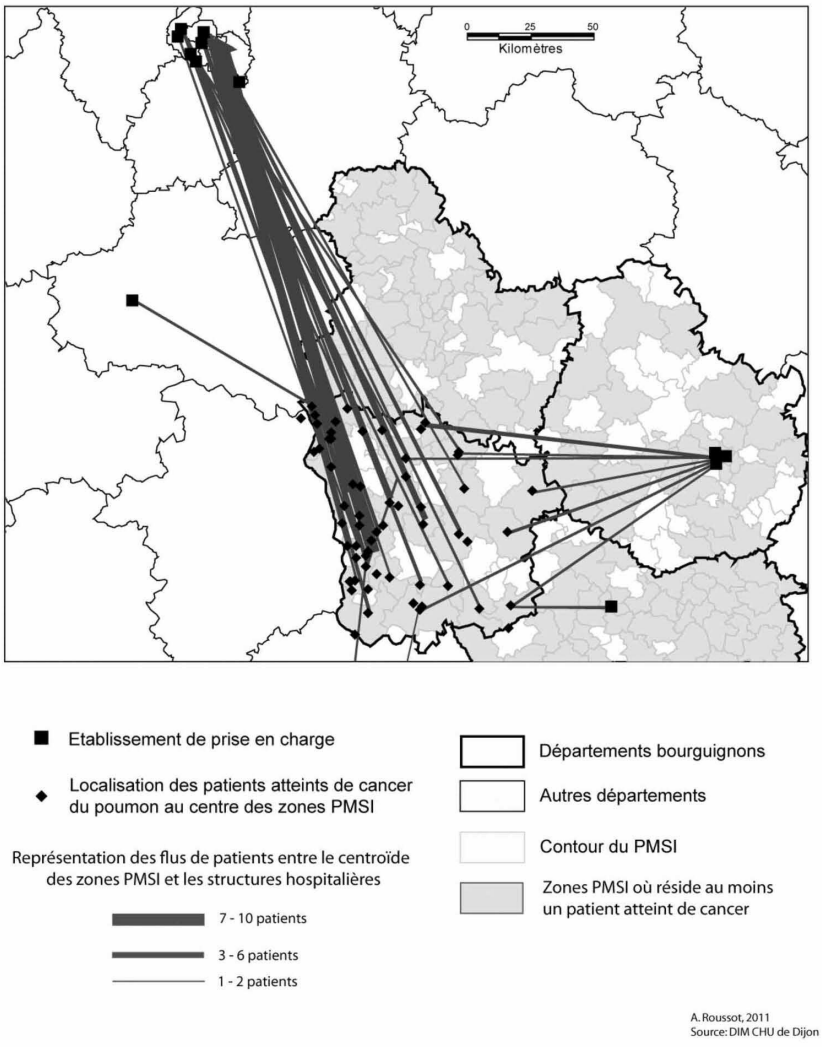


Figure 2

Prise en charge chirurgicale des cancers du poumon : patients de la Nièvre



Les trajectoires suivent des logiques méridiennes pour des patients dont on aurait pu attendre des déplacements plus nombreux vers les régions Centre et Auvergne. Il existe une dichotomie entre les territoires nivernais morvandiaux, historiquement liés à la Bourgogne des Ducs, et la Nièvre ligérienne, plus tournée vers le nord et son interface occidentale. Cette opposition entre le Morvan et la frange occidentale nivernaise ressort plus si l'on s'intéresse à la cartographie différenciée par sexe puisqu'aucune femme domiciliée dans le Morvan ou la Nièvre centrale n'a connu d'intervention chirurgicale inaugurale en Bourgogne.

L'importance des fuites pour le premier acte de chirurgie s'explique autant par l'importance du rayonnement parisien que par les carences du maillage sanitaire nivernais. Peu de spécialistes et discontinuité du réseau hospitalier, ces facteurs associés à un peuplement diffus de territoires ruraux expliquent l'absence de trajectoires intra-nivernaises lors de la première intervention chirurgicale. Les mêmes déterminants sont également valables pour les patients icaunais.

Sans distinction de sexe, les patients de l'Yonne choisissent majoritairement deux destinations de prise en charge chirurgicale : l'Ile-de-France et le CHU de Dijon (cf. Figure 3). Dans la région parisienne, les établissements fréquentés sont majoritairement des cliniques participant au service public hospitalier (PSPH) : Marie Lannelongue, 30 patients, hôpital Foch de Suresnes, 18 patients. L'Assistance Publique – Hôpitaux de Paris a reçu quant à elle 15 patients. L'axe Nord-ouest / Sud-est qui se dessine suit le tracé des grandes voies de communication et révèle l'opposition entre les attractions parisienne et dijonnaise qui tiraillent ce département de « passage ».

L'attraction dijonnaise s'observe à différentes échelles. Si elle dépasse le seul département de la Côte-d'Or et s'étend jusqu'au Morvan, elle est également particulièrement prégnante dans son environnement proche. Le premier recours chirurgical des patients côte-d'oriens coïncide ainsi avec une prise en charge au CHU de Dijon, ou, dans une moindre mesure, dans une clinique de la périphérie de Dijon, à Talant ou Chenôve (cf. Figure 4). L'attraction du pôle dijonnais suit un modèle en étoile, le centre hospitalier de Beaune étant le seul autre établissement du département à avoir reçu et opéré un patient atteint de cancer du poumon.

Les trajectoires de premier recours des patients de Saône-et-Loire sont moins structurées (cf. Figure 5). Les fuites sont importantes, mais restent majoritairement dirigées vers Lyon. On note l'importance de trois grands pôles de destination. Les patients de la frange septentrionale du département se rendent au CHU de Dijon, tandis que ceux du sud se tournent vers la région lyonnaise. Au centre du département, Chalon-sur-Saône et sa clinique exercent un tropisme de proximité.

La représentation graphique de l'analyse exploratoire (cf. Figure 6) montre que les plus grandes distances parcourues concernent les patients icaunais et nivernais qui se rendent en Ile-de-France. On voit également que les patients de Saône-et-Loire Sud sont plus enclins à effectuer leur chirurgie à Lyon ou dans un établissement de ce département, surtout du type centre hospitalier, comme ceux de Châlons-sur-Saône ou de Paray-le-Monial.

A l'inverse, les modalités « CHU/CHR » et « CL » de la variable « Type de structure » sont proches du centre du graphique et ne contribuent donc pas à la construction des axes factoriels. Celles-ci ne peuvent donc être interprétées correctement.

Classification issue Du logiciel SYR

La cartographie d'une autre méthode de classification, celle des nuées dynamiques, contribue à segmenter la patientèle étudiée sans se limiter au département d'origine pour différencier les classes de patients.

Figure 3

Prise en charge chirurgicale des cancers du poumon : patients de l'Yonne

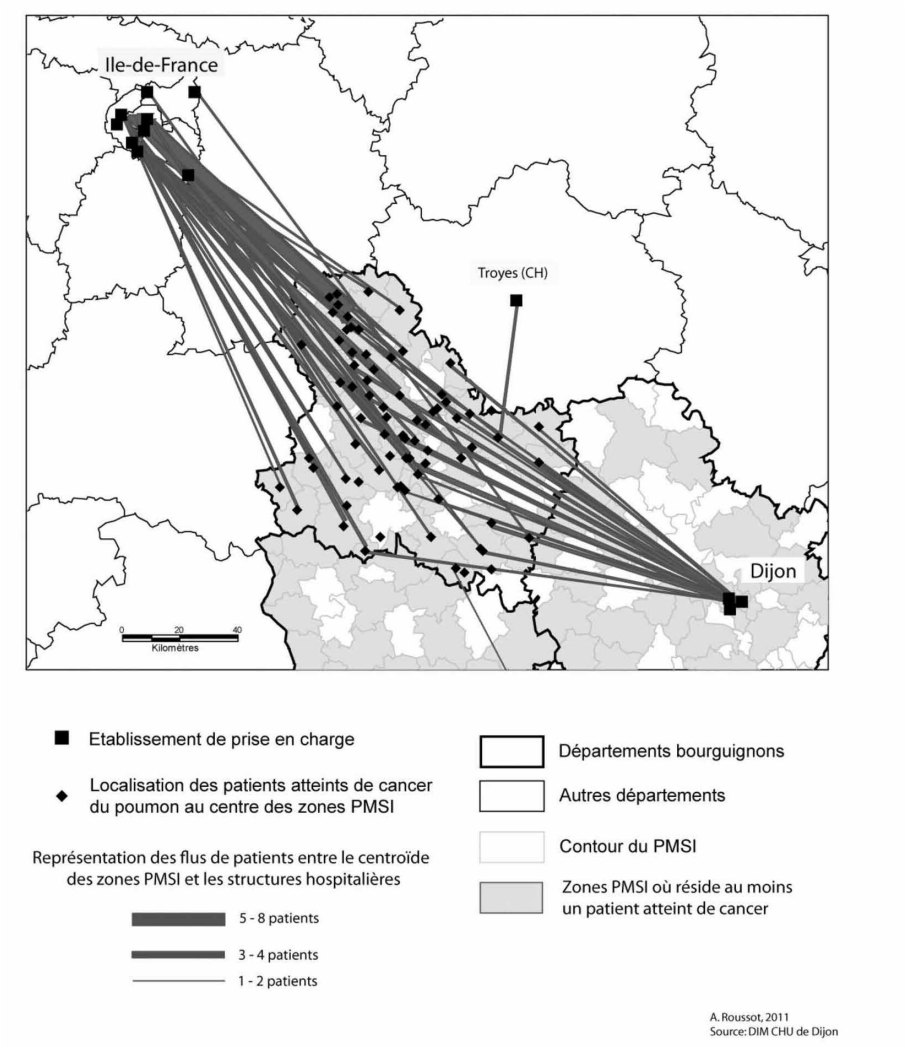


Figure 4

Prise en charge chirurgicale des cancers du poumon : patients de Côte-d'Or

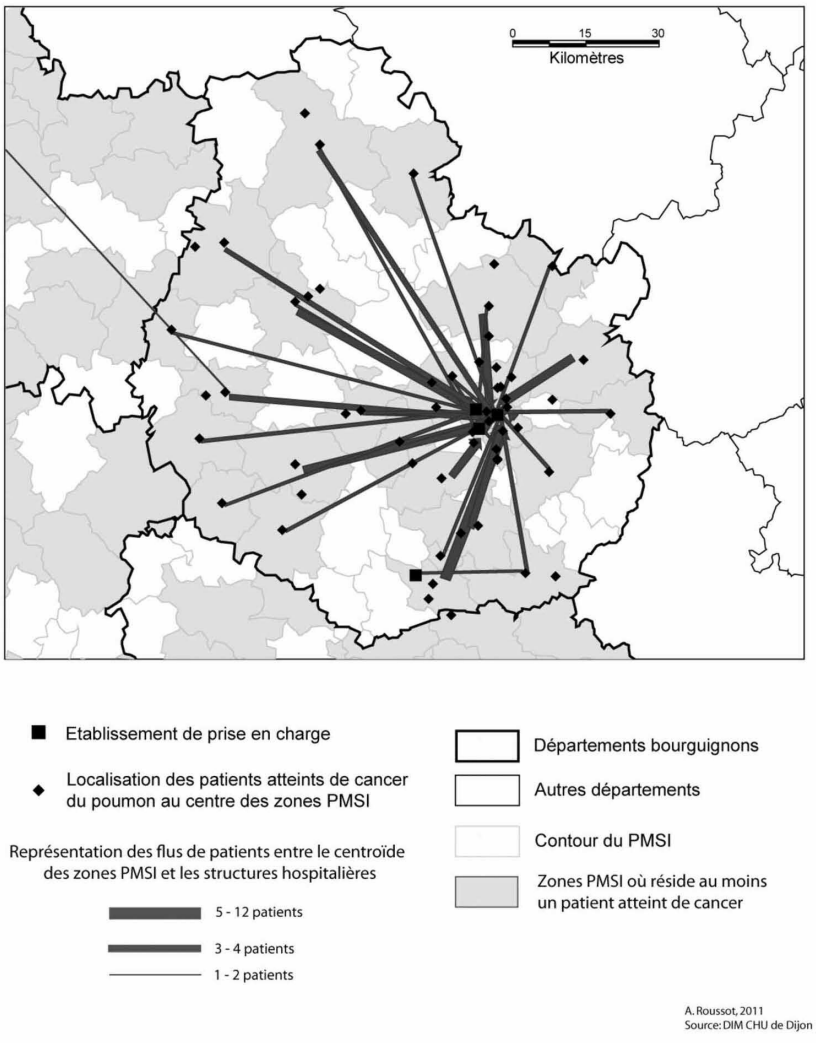


Figure 5

Prise en charge chirurgicale des cancers du poumon : patients de Saône-et-Loire

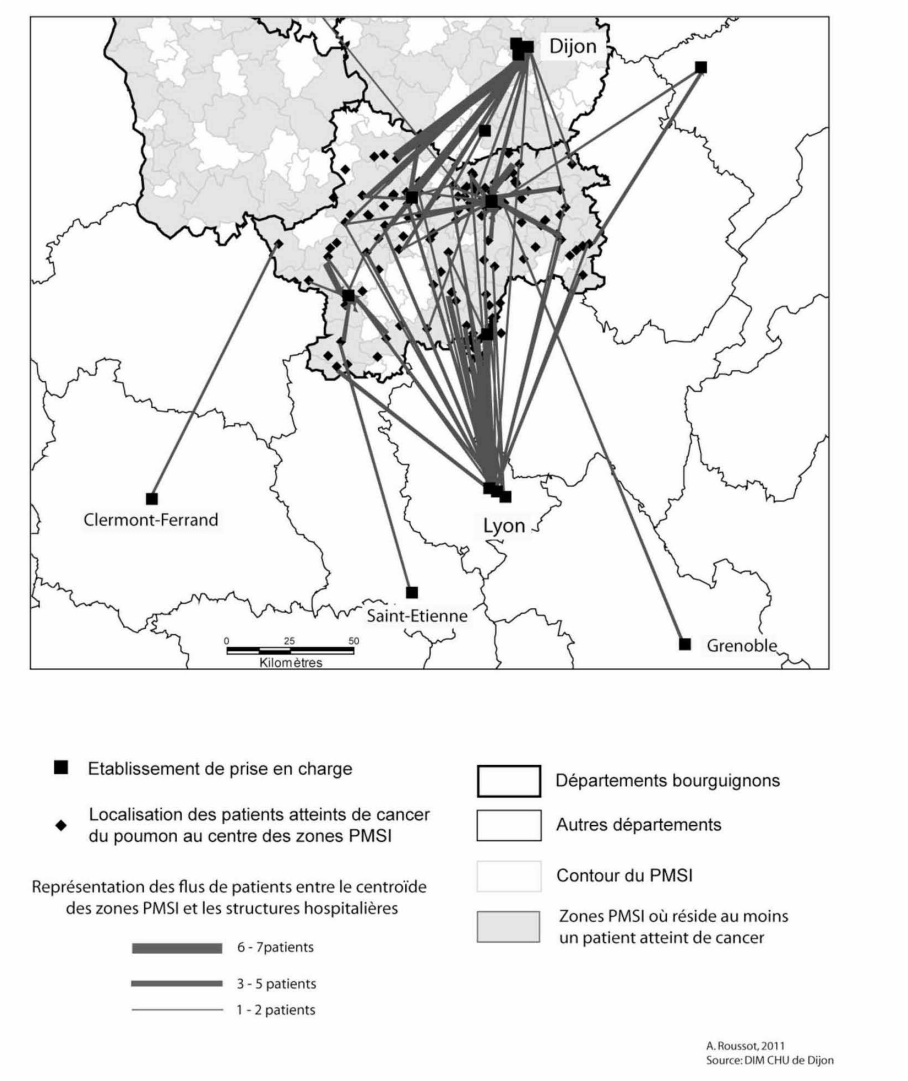
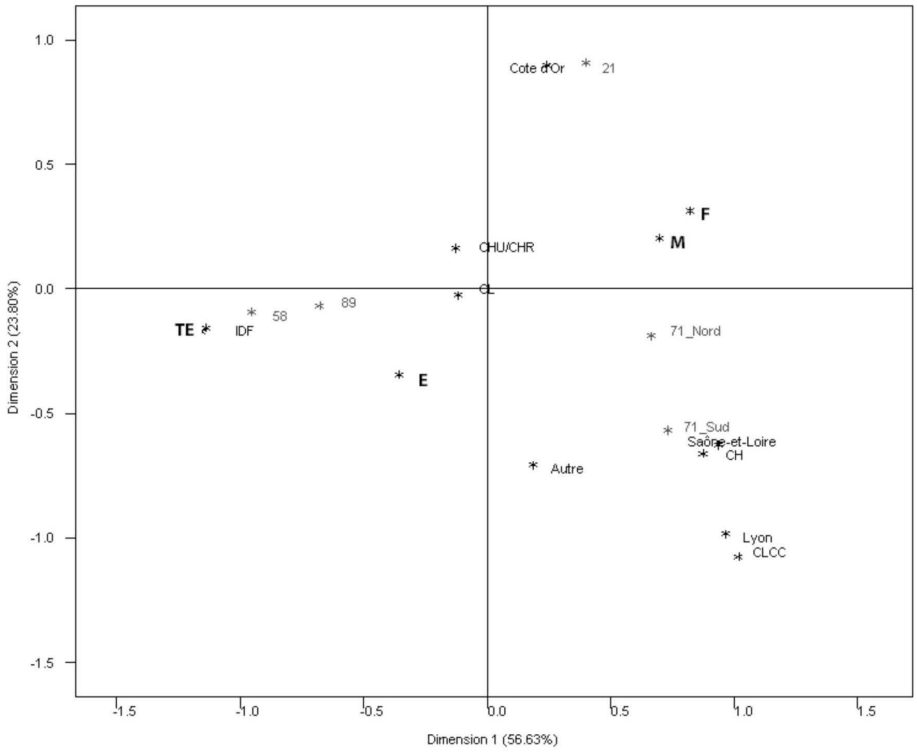


Figure 6

Relations entre le département des patients, la destination de prise en charge et la distance parcourue



Origine des patients
21 : patients résidant en Côte-d'Or
58 : patients résidant dans la Nièvre
71 : patients résidant en Saône-et-Loire
89 : patients résidant dans l'Yonne

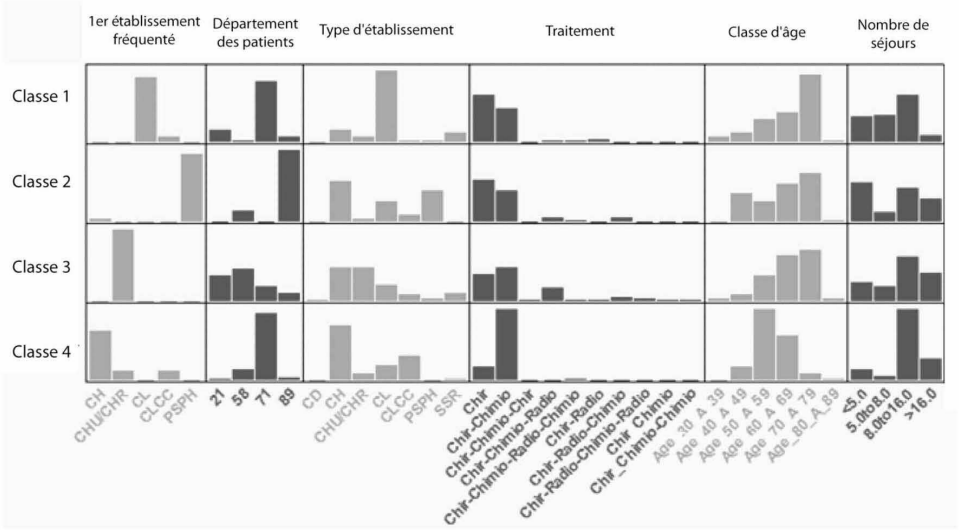
Type de structure
CHU/CHR : Centre hospitalier universitaire / Régional
CH : Centre hospitalier
CL : Clinique
CLCC : Centre de lutte contre le cancer

Distance parcourue (quartiles)
TE : très élevée : plus de 141,9 km)
E : élevée : entre 85,9 et 141,9 km
M : modérée : entre 22,3 et 85,9 km
F : faible : moins de 22,3 km

Destination de prise en charge
IDF : Ile-de-France
Lyon : Lyon et sa périphérie
Autre : autre localisation

Figure 7

Descriptif des classes issues de la classification des trajectoires.



Type de structure
CHU/CHR : Centre hospitalier universitaire / Régional
CH : Centre hospitalier
CL : Clinique
CLCC : Centre de lutte contre le cancer
PSPH : Participant au service public hospitalier

Traitement
Chir : acte de chirurgie
Chimio : séance de chimiothérapie
Radio : séance de radiothérapie

Comme le montre l'histogramme (cf. Figure 7), le premier établissement fréquenté et le département d'origine des patients sont les variables les plus discriminantes, ce qui est corroboré par la cartographie associée de la classification et des oursins de la première prise en charge. Ainsi, les patients de la classe 1 résident essentiellement en Saône-et-Loire et subissent plus souvent leur première intervention chirurgicale dans une clinique.

Les patients de la classe 2 sont surtout des Icaunais et fréquentent plus souvent des établissements privés participant au service public hospitalier (PSPH), en région parisienne. L'attraction exercée par les établissements franciliens est particulièrement prégnante pour les patients de cette classe, ce qui rejoint la cartographie des trajectoires des patients du département de l'Yonne.

La fréquentation des CHU/CHR pour une première prise en charge chirurgicale concerne les patients de la troisième classe, résidant dans les quatre départements bourguignons. Les trajectoires suivent les logiques des interfaces de la région, le rayonnement du CHU de Dijon étant plus marqué vers les territoires situés à

proximité de la Côte-d'Or. Surtout, les fuites vers l'est et l'ouest sont très limitées, la direction des trajectoires hors Bourgogne suivant une orientation méridienne.

La plupart des patients de la classe 4 résident en Saône-et-Loire, département marqué par l'importance des recours dans les centres hospitaliers (CH), mais également par le CLCC de Lyon. Comme pour les patients de la première classe, la représentation de la fréquentation des établissements suit une logique concentrique autour des CH de Paray-le-Monial et de Chalon-sur-Saône.

6. DISCUSSION

Les bases PMSI sont un matériel précieux pour l'étude des trajectoires des patients et la mise en lumière des dynamiques interterritoriales engendrées par le recours aux soins [5]. Même si le zonage proche du code postal est plus grossier que d'autres échelles administratives, celui-ci demeure un bon niveau d'analyse, qui permet de faire ressortir des dynamiques spatiales et des bassins sanitaires attractifs. Certaines logiques territoriales ressortent, et les flux de patients qui découlent des hospitalisations ne sont pas aberrants, même si leur interprétation doit tenir compte de plusieurs biais dont le principal est l'identification des établissements par leur numéro FINESS juridique. Ainsi, tous les établissements de l'AP-HP sont localisés à la même adresse et ne peuvent être individualisés.

Le travail de cartographie est destiné à apporter une représentation graphique des trajectoires, il se base essentiellement sur les parcours effectués pour le premier recours chirurgical, qui marque le début des trajectoires des patients. Le deuxième recours engage un recentrage de la trajectoire d'un patient sur les territoires de proximité, à l'instar des recours suivants. Si ces derniers n'engagent pas forcément la fréquentation d'un unique centre de soins, pour suivre une chimiothérapie par exemple, on suppose que la logique de la proximité domine pour la majorité des trajectoires des patients. Les distances parcourues pour suivre une séance de chimiothérapie sont beaucoup moins importantes que pour vivre une intervention chirurgicale, non seulement parce que les enjeux ne sont pas les mêmes, mais surtout parce que l'offre de soins pour chacune de ces prises en charge n'est pas comparable. Le hiatus entre les « trajets » de premier et de deuxième recours correspond donc pour beaucoup de patients à un retour au sein de la région pour poursuivre leur trajectoire thérapeutique, ce qui se traduit spatialement par des recours plus courts, souvent vers les petites structures.

L'analyse des données symboliques par la méthode des nuées dynamiques permet de corroborer certaines conclusions issues de l'analyse spatiale. Si les résultats cartographiques permettent de déceler des grandes tendances de trajectoires, ainsi que des grandes classes de patients, l'analyse exploratoire donne une idée de l'importance de la distance et permet de rapprocher le lieu de résidence du type d'établissement fréquenté.

Si les trajectoires suivent les logiques inhérentes à des habitudes de déplacement, elles sont néanmoins conditionnées par la disponibilité de l'offre de soins et les recommandations des spécialistes qui orientent les patients lors de leur prise en charge. Le lien entre les territoires de résidence et le lieu de recours médical, matérialisé par des flux et d'une certaine façon déterminé par les interfaces intraterritoriales, façonne au final un système de recours singulier. L'analyse spatiale des données extraites du PMSI peut rendre compte d'une organisation factuelle du recours médico-chirurgical. Cela correspond à un ensemble de trajectoires individuelles qui se sont affranchies des multiples découpages institutionnels et sanitaires et qui s'ancrent dans des territoires spécifiques mais ouverts. L'importance des flux et l'orientation des trajectoires sont autant

d'observations utiles pour l'élaboration d'une gouvernance sanitaire pragmatique.

7. CONCLUSION

Les outils d'analyse spatiale du géographe permettent de synthétiser observation locale et organisation générale du système de soins. On peut faire l'hypothèse que la méthodologie mise en œuvre pour cette analyse de la prise en charge de patients atteints de cancer du poumon peut être utilisée pour d'autres types de cancer et/ou d'autres pathologies.

Une bonne connaissance de la localisation de l'offre sanitaire fréquentée, les volumes de patientèle et des distances parcourues par les malades sont des éléments qui intéressent le planificateur parce qu'on ne peut pas planifier sans la connaissance globale du fonctionnement d'un territoire et de ses dynamiques. Le PMSI apparaît comme un outil indispensable à une bonne description de l'activité des structures, et si beaucoup considèrent l'organisation des soins comme une science, l'utilisation du PMSI participe de ce constat de D. I. Mendeleev : « la science commence là où commence la mesure ; une science exacte sans mesure serait inconcevable. »

8. REMERCIEMENTS

Ce travail s'inscrit dans le cadre de l'« Étude des trajectoires de prise en charge des cancers dans la région Bourgogne (Traj_Can) » financée par l'Institut National du Cancer (INCa).

9. RÉFÉRENCES

[1] Ministère de l'emploi et de la solidarité, mission PMSI E/3 DHOS. *Analyse des trajectoires de soins en psychiatrie*. Janvier 2001.

[2] Vigneron E. et al., *Pour une approche territoriale de la santé*. Col. Bibliothèque des territoires, DATAR/Aube. 2003.

[3] G. Nuemi, F. Afonso, C. Toque, M. Touati, E. Diday, C. Quantin. *Symbolic Data Analysis of Cancer Care Trajectories in the region of Burgundy: Application to Lung Cancers*. Workshop in Symbolic Data Analysis, Namur, Belgium. 2011.

[4] Quantin C. et al., Classification and Regression Trees on Aggregate Data Modeling: An Application in Acute Myocardial Infarction. *Journal of Probability and Statistic*. Article ID 523937.2011.

[5] Boinot L. et al., Trajectoires hospitalières des patientes atteintes de cancer du sein en Poitou-Charentes. *Revue d'épidémiologie et de santé publique*. 2007 ; 142-148.

An approach for mining care trajectories for chronic diseases

Elias Egho[1],Nicolas Jay[1], Chedy Raïssi[2], Gilles Nuemi[3], Catherine Quantin[3], Amedeo Napoli[1]

[1] Orpailleur Team, LORIA, Vandoeuvre-les-Nancy, France
{firstname.lastname}@loria.fr
[2] INRIA, Nancy Grand Est, France
{firstname.lastname}@inria.fr
[3] Service de Biostatistique et d'Information Médicale , CHU de Dijon, Dijon, France
{firstname.lastname}@chu-dijon.fr

Abstract. With the increasing burden of chronic illnesses, administrative health care databases hold valuable information that could be used to monitor and assess the processes shaping the trajectory of care of chronic patients. In this context, temporal data mining methods are promising tools, though lacking flexibility in addressing the complex nature of medical events. Here, we present a new algorithm able to extract patient trajectory patterns with different levels of granularity by relying on external taxonomies. We show the interest of our approach with the analysis of trajectories of care for colorectal cancer using data from the French casemix information system.

Keywords: datamining, chronic illness, claim data, sequential pattern mining, trajectory of care

1 Introduction

Chronic illnesses are a major burden in both developed and developing countries[5]. Patients with chronic conditions use more services and a greater array of services than other consumers. Multiple encounters of chronic patients with the healthcare system define a so-called "trajectory" of care . Lack of coordination along the trajectory of care, bad implementation of guidelines or inappropriate organization of the healthcare system may have a negative impact on quality and costs of care.

Due to the fragmentation of clinical information systems, little knowledge is readily available to describe and assess the actual processes involved in long-term care, especially in the scope of a cross-institutional analysis. However, in many countries, health information systems routinely collect medical and administrative data at regional or national scale. Among them, case-mix information systems were originally built for hospital activity report and billing purpose[2]. They hold valuable information that could help health care managers and professionals to develop inter-organizational knowledge and bring deeper insights

into inpatient care trajectories. In order to produce the expected knowledge and support decision making, case-mix systems have to be turned into longitudinal and patient-centered information systems. This requires the linkage of different stays of a same patient into a sequence that will be further processed. Because of the complex nature and extreme diversity of medical problems, patient care trajectories must be summarized and categorized for allowing meaningful inference about outcomes of particular interest.

Data-mining methods are especially adapted to the analysis of sequences and successfully used in biomedical domain [3, 4, 7, 1]. case-mix systems capture medical problems, procedures, demographic and administrative data using controlled vocabularies and standardized records. In that context, sequences of hospitalizations can be analyzed with sequential pattern mining algorithms[9]. Meanwhile, case-mix records have a multidimensional structure that traditional sequential patterns can not fully reflect. Moreover, the granularity of the initial data may be too fine to generate interesting patterns. The availability of classifications and ontologies used to code information in case-mix systems is an opportunity to integrate additional knowledge into the mining process and achieve better results. Although a few approaches have been developed to tackle the problems of granularity and multidimensionality in sequential pattern mining[8], they are still not adapted to the problem of mining care trajectories.

In this paper, we present an new algorithm, MMISP (Mining Multidimensional Itemsets Sequential Patterns). MMISP is able to extract patterns from care trajectories in a multidimensional temporal database, using external taxonomic knowledge at appropriate levels of granularity. We illustrate this approach in analysing care trajectories for colorectal cancer using data from the french case-mix information system.

2 Problem Statement

The PMSI[4] is the french adaptation of the Diagnoses Related Groups[2]. In the PMSI database, each stay is a standardized record of administrative and clinical data, especially about the institution, the patient's principal diagnosis and the realized medical procedures. In order to formalize the problem, we accordingly model each hospitalization along three dimensions: (*i*) healthcare institution, (*ii*) diagnosis and (*iii*) medical procedure. Two dimensions, i.e. healthcare institutions and diagnosis, are considered as ordered sets with an associated subsumption relation (i.e. a partial ordering). The set of healthcare institutions H, the set of diagnosis DG and the set of medical procedures MP, are given:

- H=$\{t_h, uh, gh, uh_p, uh_n, gh_p, gh_l\}$.
- DG=$\{t_d, c, r, c_1, c_2, r_1, r_2\}$.
- MP=$\{mp_1, mp_2, mp_3, mp_4\}$.

The subsumption relation for H and DG is defined as below (Figure 1).

[4] Programme de Médicalisation des Sytèmes d'Information

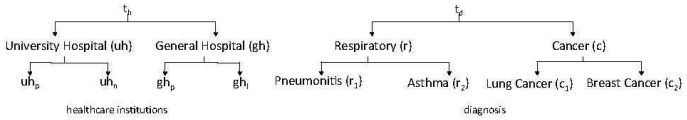


Fig. 1. The subsumption relation for the healthcare institutions set H and the diagnosis set DG

Definition 1. *A* **partially ordered set (poset)** *is a pair $(D, \leqslant)$, where D is a set and $\leqslant$ is a partial order relation on D. For $x \in D$ the* **down set** *of x, denoted by $\downarrow x$, is a set of all specializations of x; $\downarrow x = \{y \in D | y \leqslant x\}$. The* **up-set** *of x is $\uparrow x = \{y \in D | x \leqslant y\}$.*

Among the three basic dimension, H and DG are posets. The hospitalization of a patient is then considered as a vector with 3 components, (H, DG, MP).

Example 1. $(uh_p, c_1, \{mp_1, mp_2\})$ is an hospitalisation for a patient. It is a vector with three components $uh_p \in H, c_1 \in DG$ and $\{mp_1, mp_2\} \subseteq MP$.

Definition 2. *(Elementary vector) An elementary vector $v = (v_1, v_2, v_3)$ is a vector with 3 elements. Given two vectors $v = (v_1, v_2, v_3)$ and $v' = (v_1', v_2', v_3')$, v is more general than v', denoted by $v' \leq_v v$, for every $i = 1...3$*

$v_i' \leqslant v_i$ *if v_i, v_i' are elements in a poset*
$v_i \subseteq v_i'$ *if v_i, v_i' are sets*

Example 2. $v = (uh_p, c_1, \{mp_1, mp_2\})$ is a vector with 3 elements uh_p, c_1 and $\{mp_1, mp_2\}$. The vector $v' = (uh, t_d, \{mp_1\})$ is more general than v, $v \leq_v v'$, because of:

- $uh_p \leqslant uh$; $uh_p, uh \in H$
- $c_1 \leqslant t_d$; $c_1, t_d \in DG$.
- $\{mp_1\} \subseteq \{mp_1, mp_2\}$; $\{mp_1\}, \{mp_1, mp_2\} \subseteq MP$.

Definition 3. *(Patient Trajectory) A patient trajectory is a pair (V,$<_t$), where V is a set of elementary vectors and $<_t$ is a temporal order relation on V. The patient trajectory represents like $P = \langle P_1 P_2 ... P_l \rangle$, where $P_1, P_2, ..., P_l \in V$ and $P_1 <_t P_2 <_t P_3 ... <_t P_l$. Given two trajectories $P = \langle P_1 P_2 ... P_l \rangle$ and $T = \langle T_1 T_2 ... T_{l'} \rangle$, P is more general than T, denoted by $T \leq_p P$, if there exist indices $1 \leq i_1 < i_2 < ... < i_l \leq l'$ such that $T_j \leq_v P_{i_j}$ for all $j = 1 \ldots l$ and $l \leqslant l'$. We say that T is more specific than P.*

Example 3. $\langle (uh_p, c_1, \{mp_1, mp_2\})(gh_l, r_1, \{mp_2\}) \rangle$ represents a patient trajectory with two hospitalizations. It expresses the fact that a patient was admitted to the hospital uh_p for a lung cancer c_1, and underwent procedures mp_1 and mp_2. Then he went to the hospital gh_l for pneumonitis r_1 where he underwent procedure mp_2.

Patients	Trajectories
$patient_1$	$\langle(uh_p, c_1, \{mp_1, mp_2\})(uh_p, c_1, \{mp_1\})(gh_l, r_1, \{mp_3\})\rangle$
$patient_2$	$\langle(uh_n, c_1, \{mp_4\})(uh_p, c_2, \{mp_1, mp_2\})(gh_l, r_1, \{mp_2\})\rangle$
$patient_3$	$\langle(uh_p, c_1, \{mp_4\})(gh_l, r_2, \{mp_3\})\rangle$
$patient_4$	$\langle(uh_p, c_2, \{mp_1, mp_2\})(gh_p, r_2, \{mp_3\})(gh_l, r_2, \{mp_2\})\rangle$

Table 1. An example of a database of patient trajectories.

Let P_{DB} be the patient trajectories for four patients $patient_1$, $patient_2$, $patient_3$ and $patient_4$, Table 1.

Let `supp(P)` be the number of trajectories that are more specific than P in P_{DB} and σ be a minimum support threshold specified by the end-user. Let P be a trajectory, P is a frequent trajectory pattern in P_{DB} if and only if $supp(P) \geq \sigma$.

Using the poset for some dimensions, we can extract a large number of frequent trajectory patterns. To avoid the patterns overfloading, our approach only extracts the set of all most specific frequent trajectory patterns in P_{DB}. Actually, frequency is anti-monotonic (i.e if $P = \langle(uh_p, c1, \{mp_1, mp_2\})\rangle$ is a frequent then $T = \langle(uh, c, \{mp_1\})\rangle$ which is more general than P is also frequent). So, all the most specific frequent trajectory patterns can lead to some general frequent trajectory patterns.

Definition 4. *(Most Specific Frequent Trajectory) Let P be a trajectory. P is a most specific frequent trajectory, if and only if: $supp(P) \geq \sigma$ and for all T such that $T \leq_p S$; $supp(T) \leqslant \sigma$*

Example 4. Let $\sigma = 0.75$ (i.e. a trajectory is frequent if it appears at least three times in P_{DB}). The trajectory $P = \langle(uh_p, c, \{mp_1, mp_2\})\rangle$ is frequent. $T = \langle(uh, c, \{mp_1\})\rangle$ is also frequent. Nevertheless, T is not a most specific frequent trajectory pattern while P is one.

3 Mining patient trajectory patterns

In this section, we present an approach for extracting all the most specific frequent trajectory patterns from patients trajectories. Our approach is called **MMISP** (*Mining Multidimensional Itemsets Sequential Patterns*). The basic idea of MMISP is finding a way to transfer the multidimensional itemsets sequential database into a classical sequential database (i.e. sequence of itemsets). So, MMISP is based on three steps:

1. Extract all the frequent elementary vector v without taking into account the temporal relation between them in each trajectory.
2. Map the frequent elementary vectors which extracted in the first step to an alternate representation. Then, the patient trajectories are encoded by using the new representation of frequent elementary vectors.
3. Apply a standard sequential mining algorithm to enumerate frequent patient trajectories.

3.1 Generating frequent elementary vectors

MMISP starts by searching for the frequent elementary vectors in the trajectories. MMISP firstly studies the patient's trajectory like a set of elementary vectors without taking into account the temporal relation order between them. The support of elementary vector v is defined as follows,

Definition 5. *(Support of elementary vector v, supp(v)) Let P_{DB} be a database of patient trajectories with m patients and let $P =< P_1 P_2 ... P_l >$ be a patient trajectory in P_{DB}. The support of elementary vector v is defined as follows*

$$supp(v) = \frac{|\{P \in P_{DB}; \exists j \in [1,..,l];\ P_j \leq_v v\}|}{m}$$

Example 5. In our example, the support of $(gh, r, \{mp_3\})$ is $\frac{3}{4}$, because of:

- $(gh_l, r_1, \{mp_3\}) \in patient_1$ where $(gh_l, r_1, \{mp_3\}) \leq_v (gh, r, \{mp_3\})$.
- $(gh_l, r_2, \{mp_3\}) \in patient_3$ where $(gh_l, r_2, \{mp_3\}) \leq_v (gh, r, \{mp_3\})$.
- $(gh_p, r_2, \{mp_3\}) \in patient_4$ where $(gh_p, r_2, \{mp_3\}) \leq_v (gh, r, \{mp_3\})$.

MMISP generates all the frequent elementary vectors by building a poset $(L, \leq_v)$. Building $(L, \leq_v)$ is done as follows:

- Firstly, we generate the most general elementary vector. In our running example, we have two dimensions with posets H and DG and one dimension with a set MP, so the most general elementary vector is $(t_h, t_d, \{\})$.
- Then, the recursive generation of the new elementary vectors continues by using each previously generated frequent elementary vector (v). For each element $v_1, v_2, v_3 \in v$, we replace v_k, where $k \in [1,3]$ with each of its specialization from the set $special(v_k)$. At each step, we take only the frequent elementary vector which has support greater than σ.

We define the set $special(v_i)$ as follows:

Definition 6. *Let v_i be the i^{th}-element in the vector $v = (v_1, v_2, v_3)$ and let D be the ground set of the component v_i*

$$special(v_i) = \begin{cases} \{a \in D; a \leq v_i \text{ and } \nexists b \in D;\ a \leq b \text{ and } b \leq v_i\} & \text{if } D \text{ is a poset} \\ \{v_i \cup \{a\}; a \in D \setminus v_i\} & \text{if } D \text{ is a set} \end{cases}$$

Example 6. In our example, $special(t_h)$=$\{uh, gh\}$, $special(t_d)$=$\{r, c\}$ and $special(\{\})$ =$\{\{mp_1\}, \{mp_2\}, \{mp_3\}, \{mp_4\}\}$. With $\sigma = \frac{3}{4}$ we can generate new seven frequent elementary vectors from $(t_h, t_d, \{\})$. They are $(uh, t_d, \{\})$, $(gh, t_d, \{\})$, $(t_h, r, \{\})$, $(t_h, c, \{\})$, $(t_h, t_d, \{mp_1\})$,$(t_h, t_d, \{mp_2\})$ and $(t_h, t_d, \{mp_3\})$. The first and the second are generated by replacing t_h by $child(t_h)$, the third and the forth are generated by replacing t_d by $special(t_d)$, and the rest are generated by replacing $\{\}$ by $special(\{\})$.

The objective of MMISP is to generate all the most specific frequent patient trajectories, thus it retains only the most specific frequent elementary vectors from $(L, \leq_v)$.

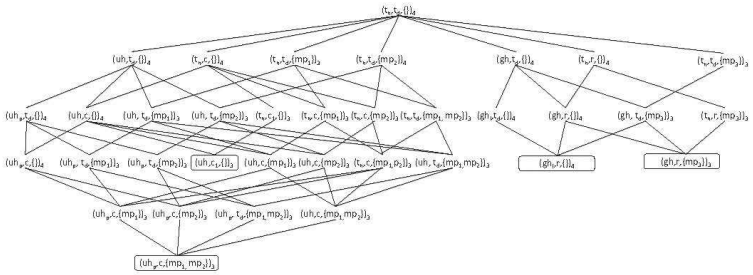


Fig. 2. The poset $(L, \leq_v)$ is generated by taking into account the two posets H and DG in Figure 1 and the set $MP = \{mp_1, mp_2, mp_3, mp_4\}$ with minsup= $\frac{3}{4}$.

Definition 7. *(Most specific frequent elementary vector, MSFV) Let v be an elementary vector, v is a most specific frequent elementary vector, if and only if $supp(v) \geq \sigma$ and $\nexists\ v'$ an elementary vector, where $supp(v) = supp(v')$ and $v' \leqslant_v v$.*

id	*MSFV*
1	$(uh_p, c, \{mp_1, mp_2\})$
2	$(uh, c_1, \{\})$
3	$(gh_l, r, \{\})$
4	$(gh, r, \{mp_3\})$

Table 2. The most specific frequent elementary vectors extracted from $(L, \leq_v)$ in Figure 2.

Example 7. Figure 2 illustrates the generation of all frequent elementary vectors on our example with $\sigma = \frac{3}{4}$. Table 2 shows the hash table of all *MSFV* extracted from $(L, \leq_v)$.

3.2 Mining patient trajectory

The next step of MMISP is studying the temporal relation between the most specific frequent elementary vectors extracted in previously step. This is done by taking each patient trajectory $P = \langle P_1 P_2, ..., P_l \rangle$ from the database of patient trajectories P_{DB}, then replacing each elementary vector $P_i \in P;\ i \in [1..l]$ with all elementary vectors $v \in MSFV$ where $P_i \leq_v v$.

Example 8. In our example, the trajectory of $patient_3$,$\langle(uh_p, c_1, \{mp_4\})(gh_l, r_2, \{mp_3\})\rangle$, is transformed into $\langle\{(uh, c_1, \{\})\}\{(gh_l, r, \{\}), (gh, r, \{mp_3\})\}\rangle$ because the first elementary vector of $patient_3$, $(uh_p, c_1, \{p_4\})$, can only be replaced with $(uh, c_1, \{\})$ from the $MSFV$ set where $(uh_P, c_1, \{p_4\}) \leqslant_v (uh, c_1, \{\})$ and the second elementary vector of $patient_3$, $(gh_l, r_2, \{mp_3\})$, can be replaced by $(gh_l, r, \{\})$ and $(gh, r, \{mp_3\})$ from the $MSFV$ set.

Table 3 shows the transformation of patient trajectories in P_{DB} by using the set of all most specific frequent elementary vector $MSFV$ in Table 2.

Patients	Trajectories
$patient_1$	$\langle\{(uh_p, c, \{mp_1, mp_2\}), (uh, c_1, \{\})\}\{(uh, c_1, \{\})\}\{(gh_l, r, \{\}), (gh, r, \{mp_3\})\}\rangle$
$patient_2$	$\langle\{(uh, c_1, \{\})\}\{(uh_p, c, \{mp_1, mp_2\})\}\{(gh_l, r, \{\})\}\rangle$
$patient_3$	$\langle\{(uh, c_1, \{\})\}\{(gh_l, r, \{\}), (gh, r, \{mp_3\})\}\rangle$
$patient_4$	$\langle\{(uh_p, c, \{mp_1, mp_2\})\}\{(gh, r, \{mp_3\})\}\{(gh_l, r, \{\})\}\rangle$

Table 3. Transforming a patient trajectories in Table 1 by using the set of all most specific frequent elementary vector in Table 2.

We apply a classical sequential pattern mining algorithm (e.g. [6, 11, 10]) to extract the frequent sequential patterns. This extraction has be done as follows: firstly we transform each patient trajectory into a sequence simple (i.e sequence of itemset like $< \{a, b\}\{a, d\} >$) and then we apply a CloSpan [10] on the transformation patient trajectories. The transformation has be done as follows:

- Each elementary vector in the $MSFV$ set is assigned a unique id which will be used during the mining operation. This is illustrated in Table 2 .
- For each elementary vector v in a patient trajectory in Table 3, we replace v with its id in Table 2.

Example 9. In our example, the patient trajectory $patient_3 = \langle\{(uh, c_1, \{\})\}\{(gh_l, r, \{\}), (gh, r, \{mp_3\})\}\rangle$ in Table 3 is transformed into $\langle\{2\}, \{3, 4\}\rangle$, because $(uh, c_1, \{\})$ has an id 2, $(gh_l, r, \{\})$ has an id 3 and $(gh, r, \{mp_3\})$ has an id 4.

Patients	Trajectories
$patient_1$	$\langle\{1, 2\}\{2\}\{3, 4\}\rangle$
$patient_2$	$\langle\{2\}\{1\}\{3\}\rangle$
$patient_3$	$\langle\{2\}\{3, 4\}\rangle$
$patient_4$	$\langle\{1\}\{4\}\{3\}\rangle$

Table 4. Transformed database in Table 3

Table 5 displays all frequent sequences in their transformed format and the frequent patient trajectories in which identifiers are replaced with their actual values with minsup=$\frac{3}{4}$.

Frequent sequential patterns	Frequent patient trajectory patterns	Support
$\langle\{3\}\rangle$	$\langle(gh_l, c)\rangle$	1
$\langle\{2\}\{3\}\rangle$	$\langle(uh, c_1)(gh_l, c)\rangle$	0.75
$\langle\{4\}\rangle$	$\langle(gh, r, \{mp_3\})\rangle$	0.75
$\langle\{1\}\{3\}\rangle$	$\langle(uh_p, c, \{mp_1, mp_2\})(gh_l, c)\rangle$	0.75

Table 5. Frequent patient trajectory patterns with minsup=$\frac{3}{4}$.

4 Results

This section describes the results obtained with MMISP on a set of 2618 trajectories of care of patients from the Burgundy region in France. Using data from the PMSI, the so-called french case mix system, we reconstituted the sequence of hospitalizations of patients having undergone surgery for colorectal cancer between 2006 and 2008, with a one year follow-up. Each event in a sequence was characterized by the following dimensions : hospital, principal diagnosis, procedures delivered during the stay. The hospital dimension was associated with a geographical taxonomy of 4 levels : root (France), administrative region, administrative department, hospital. Principal Diagnosis could be described at 5 levels of the 10^{th} International classification of Diseases (ICD10): root , chapter, block, 3-character, 4-character, terminal nodes. Procedures were represented by their first CCAM[5] code.

Figure 3 shows the number of discovered patterns at different thresholds according to their length. The total number of patterns grows exponentially for support below 34%. However, the increase is extremely variable considering the length of patterns and the number of short patterns (length<6) is still manageable. The high number of length 7 patterns can probably be explained by a combinatorial effect resulting from a high number of sequences of length 14-15 in the database. They correspond to the patients who underwent chemotherapy and usually had around 14 and 15 stays for 1 cycle.

Table 6 shows the items appearing in the Principal Diagnoses dimension of patterns for which support is over 32%. It can be noticed that the ICD10 tree has been mined at different levels. In the neoplasm branch, the most specific observed item is of depth 3, Malignant neoplasm of colon. In the branch of "Factors influencing ...", items of depth 4 (chemotherapy session for neoplasm) have been extracted. Children of "Malignant neoplasms of colon" are not frequent enough

[5] Classification Commune des Actes Médicaux : the french classification of medical and surgical procedures

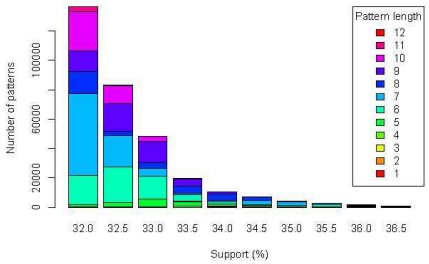


Fig. 3. Number of sequential patterns by support and length (stacked bars).

to be extracted, but "chemotherapy session" appears in a sufficient proportion of trajectories to be seen. Such results could not have been obtained by representing items at an arbitrary pre-determined level.

ICD10 level – Item
0– Root
1– Neoplasms
2– Malignant neoplasms of digestive organs
3– Malignant neoplasm of colon
1– Factors influencing health status and contact with health services
2– Persons encountering health services for specific procedures and health care
3– Other medical care
4– Chemotherapy session for neoplasm

Table 6. Items extracted in the Principal Diagnosis dimension, (minsupp=32%)

Multidimensional sequential patterns can be analysed per se. For example, the pattern $\langle$(Root,C15-C26, {Colectomy}), (Burgundy,Z00-Z99,{})$\rangle$ shows that 69% of patients had a colectomy for a digestive cancer and a subsequent stay in the Burgundy region for complementary treatments and follow-up. This kind of information can help healthcare managers and deciders in planning and organizing healthcare resources at a regional level. Besides, sequential patterns can be seen as condensed representations of the care trajectories. As such, they can be reused as new variables to distinguish subgroups of patients in subsequent analysis. As an illustrative example, we selected a subset of frequent patterns to analyze the relationship between accessibility of care facilities and trajectories of care, as it has been shown that geographical disparities might be related to less favourable outcome in terms of survival. The cumulative driving distance trav-

elled by patients to access facilities along their care trajectory was used to fit a classification tree with patterns as predictors. As expected, longer distances were associated to trajectories involving chemotherapy sessions. However differences were observed according to the occurrence of hospitalizations in specific places. In particular, patients initially treated outside of the burgundy region travelled longer distances. Theses findings can bring experts to investigate specific hypothesis regarding the links between organization of care and health outcomes.

5 Conclusion

Care trajectories of chronic patients can be analysed using administrative databases and sequential pattern mining. The MMISP algorithm relies on external knowledge to enrich the mining process and produces results with appropriate levels of granularity. Experiments on data from the french case-mix information system show that MMISP is flexible enough to reflect both the relational and temporal structure of the care trajectories.

References

1. Iyad Batal, Lucia Sacchi, Riccardo Bellazzi, and Milos Hauskrecht. A temporal abstraction framework for classifying clinical temporal data. *AMIA Annu Symp Proc*, 2009:29–33, 2009.
2. RB Fetter, Y Shin, JL Freeman, RF Averill, and JD Thompson. Case mix definition by diagnosis-related groups. *Med Care*, 18(2):1–53, Feb 1980.
3. Zhengxing Huang, Xudong Lu, and Huilong Duan. On mining clinical pathway patterns from medical behaviors. *Artif Intell Med*, 56(1):35–50, Sep 2012.
4. Mingoo Kim, Hyunjung Shin, Tae Su Chung, Je-Gun Joung, and Ju Han Kim. Extracting regulatory modules from gene expression data by sequential pattern mining. *BMC Genomics*, 12 Suppl 3:S5, Nov 2011.
5. Ellen Nolte and Martin McKee, editors. *Caring for people with chronic conditions : A health system perspective*. Open University Press, 2008.
6. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE*, pages 215–224, 2001.
7. François Petitjean, Florent Masseglia, Pierre Gançarski, and Germain Forestier. Discovering significant evolution patterns from satellite image time series. *Int J Neural Syst*, 21(6):475–489, Dec 2011.
8. Marc Plantevit, Anne Laurent, Dominique Laurent, Maguelonne Teisseire, and Yeow WEI Choong. Mining multidimensional and multilevel sequential patterns. *ACM Trans. Knowl. Discov. Data*, 4:4:1–4:37, January 2010.
9. Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK, 1996. Springer-Verlag.
10. Xifeng Yan, Jiawei Han, and Ramin Afshar. Clospan: Mining closed sequential patterns in large datasets. In *In SDM*, pages 166–177, 2003.
11. Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1-2):31–60, January 2001.

Jay et al. BMC Medical Informatics and Decision Making 2013, 13:130
http://www.biomedcentral.com/1472-6947/13/130

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer

Nicolas Jay[1,2*], Gilles Nuemi[3], Maryse Gadreau[5] and Catherine Quantin[3,4]

Abstract

Background: With the increasing burden of chronic diseases, analyzing and understanding trajectories of care is essential for efficient planning and fair allocation of resources. We propose an approach based on mining claim data to support the exploration of trajectories of care.

Methods: A clustering of trajectories of care for breast cancer was performed with Formal Concept Analysis. We exported Data from the French national casemix system, covering all inpatient admissions in the country. Patients admitted for breast cancer surgery in 2009 were selected and their trajectory of care was recomposed with all hospitalizations occuring within one year after surgery. The main diagnoses of hospitalizations were used to produce morbidity profiles. Cumulative hospital costs were computed for each profile.

Results: 57,552 patients were automatically grouped into 19 classes. The resulting profiles were clinically meaningful and economically relevant. The mean cost per trajectory was 9,600€. Severe conditions were generally associated with higher costs. The lowest costs (6,957€) were observed for patients with in situ carcinoma of the breast, the highest for patients hospitalized for palliative care (26,139€).

Conclusions: Formal Concept Analysis can be applied on claim data to produce an automatic classification of care trajectories. This flexible approach takes advantages of routinely collected data and can be used to setup cost-of-illness studies.

Keywords: Data mining, Formal concept analysis, Claim data, Trajectory of care, Cancer

Background

Health-care systems face a crisis of an increasing burden of chronic diseases aggravated by aging populations [1]. It is of much importance that policy makers and healthcare managers can make decisions based on sufficient knowledge and understanding of chronic care activities. This is especially true in the field of cancer where incidence, therapeutics, practices and costs can vary quickly [2,3]. On the one hand, policy-makers need cost-effectiveness and cost-of-illness analyzes for planning and fair allocation of funding. On the other hand, care providers should be able to adapt their resources and costs while they share patients in multidisciplinary and coordinated approaches. Costs can be estimated from a variety of data sources, including insurance claims, billing systems, hospital discharge databases and surveys [4]. However, data sources may vary in a number of important aspects: accessibility, representativeness, level of aggregation, period of observation, availability and accuracy of clinical data. Besides, discrepancies can be observed depending on the source used to identify cases or estimate medical expenditures [5].

In parallel to ad hoc surveys that are often temporary and costly, administrative data are routinely collected in perennial information systems. They are an easily accessible source of information to analyze the economical burden of chronic diseases [6]. Moreover, when they contain enough clinical details, claim databases have proven to be useful in the field of epidemiology [7-12]. Though essentially used for funding and analysis of isolated episodes

*Correspondence: nicolas.jay@univ-lorraine.fr
[1]Université de Lorraine, LORIA UMR 7503, F-54000, Nancy, France
[2]CHU de Nancy, Département d'information médicale, F-54000, Nancy, France
Full list of author information is available at the end of the article

of care, this combination of medical and economical information may contain sufficient ingredients to study trajectories of care, giving better and more comprehensive insights on the journey of chronic patients in the healthcare system [13].

In France, the *Programme de Médicalisation des Systèmes d'Information* (PMSI) is a nationwide information system, derived from the Diagnosis Related Groups (DRG) system [14]. Initially build for billing purposes, the PMSI system has two important advantages for the analysis of trajectories of care. Each sector of activity (acute, post-acute, psychiatric care) is covered by an information system common to the whole French population. Second, since the introduction of an anonymised identifier in 2001, it allows the linkage of all hospitalizations of a same patient across time, space and sectors of activity. However, as most of existing patient classification systems, the PMSI focuses on single contacts and was not designed to categorize a care process spanning several encounters. For chronic conditions, it is of much interest to summarize the information contained in a set a longitudinal data and produce meaningful categories that will be relevant for subsequent analysis. This is a difficult and time consuming classification task, as chronic patients can have multiple diagnoses and multiple treatments recorded in several different facilities, and because it is an indirect and a posteriori use of data that were initially collected for budgetary purposes. Besides, different classifications may be required to achieve different goals. Meanwhile, data mining methods may support the experts in the categorization and analysis of trajectories of care [15].

In this article, we propose a method for grouping trajectories of care over a sequence of hospitalizations, using claim data. Our approach relies upon Formal Concept Analysis (FCA), a conceptual clustering method, and data from the PMSI. We studied one-year trajectories of care of the patients having undergone breast cancer surgery in 2009 in France.

Methods

Formal concept analysis

Introduced by Wille [16], Formal Concept Analysis (FCA) is a theory of data analysis identifying conceptual structures within data sets [17]. FCA is closely related to the well-known Association Rule Mining (ARM) and frequent itemsets discovery methods [18]. Indeed, many of the most efficient ARM algorithms are FCA-based [19-21]. A key advantage of the FCA-like mining lays in the fact that due to closure properties, only patterns of maximal size are extracted. This ability to produce condensed representation of patterns or rules reduces the exploration/interpretation burden for the analyst [22]. Another strong feature of FCA is its capability of discovering inherent hierarchical structures within data and thereby producing graphical visualizations. FCA has been successfully used in various health related applications [23-28].

FCA mathematizes the philosophical understanding of a concept as a knowledge unit consisting of two parts: the extent and the intent. The extent covers all objects (or entities) that are instances of the concept, while the intent comprises all attributes (or properties) holding for all the objects under consideration. FCA starts with a formal context defined as a triple $K = (G, M, I)$ where G is a set of objects, M a set of attributes and I a binary relation between G and M. $(g, m) \in I$ means that the object g has the attribute m. K may be seen as a table relating objects and their attributes. The Table 1 shows a formal context K representing the relation I between a set of 8 objects $G = \{1, \ldots, 8\}$ and a set of 4 attributes $M = \{a, b, c, d\}$. A cross indicates that a given object has the corresponding attribute. Two operators both denoted by $'$ can be defined on the power sets of objects 2^G and attributes 2^M as follows:

$$' : 2^G \rightarrow 2^M, X' = \{m \in M | \forall g \in X, (g, m) \in I\}$$

The $'$ operator is dually defined on attributes. A formal concept of K is a pair (A, B) with $A \subseteq G$ and $B \subseteq M$ such that $A' = B$ and $B' = A$. A is called the extent and B is called the intent of the formal concept. For example, $C = (\{g2, g4\}, \{m1, m2\})$ is a formal concept of K. A subconcept - superconcept relation can be formalized as:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1$$

The set of all concepts of a formal context $K = (G, M, I)$ together with the order relation form a complete lattice and can be displayed in a line diagram as shown in the right part of Figure 1 for the formal context of Table 1. Such diagrams can be very useful in the field of knowledge discovery to understand conceptual relationships among data. However, the number of formal concepts increases, at worst exponentially, with the size of the formal context. Some interest measures for concepts have been proposed to reduce the complexity of concept lattices [29-31]. The

Table 1 A formal context *K*

	a	b	c	d
1	X			
2		X	X	
3		X		X
4		X	X	
5	X			
6	X	X	X	
7			X	
8		X		X

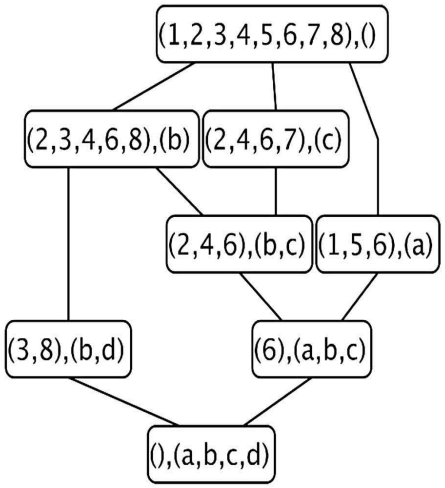


Figure 1 Concept lattice of the formal context *K*.

support of a concept (A, B) is the number of objects in its extent [29]:

$$support(A, B) = |A|$$

Among these objects, some may have exactly the properties of the intent and no more. Such objects are called the own objects of a concept. Formally, given a concept (A, B), the set of own objects of this concept is:

$$\{g \in A | \{g\}' = B\}$$

For example, in Figure 1, 7 is an own object of the concept $\{2, 4, 6, 7\}, \{c\}$. The concept $\{2, 3, 6, 8\}, \{b\}$ has no own object because 3 and 8 share also *d* and 2, 4, 6 share also *c*. The proportion of own objects in a concept is a measure of the cohesion of its objects. When this proportion is low, the objects of the concept are likely to have a more specific description and to also appear in its subconcepts. The concept may then be seen as not comprehensive enough. Noise in data tends to generate such concepts. In our experiment, we have filtered the lattice by discarding concepts with low support and low proportion of own objects.

The PMSI database

The PMSI is the French casemix system database. It is mandatory in all public and private hospitals where each admission triggers the collection of a minimal set of data holding administrative and clinical information. The PMSI database hold 24,575,239 stays in 2009. The coding of diagnoses and procedures forms the basis of information needed for the definition of patient groups. Diagnoses are coded with the 10th International Classification of Diseases (ICD-10) and medical procedures with the french nomenclature "Classification Commune des Actes Médicaux (CCAM)". In the PMSI, diagnoses can have different roles: principal, related or comorbidity. The principal diagnosis is the condition problem that is chiefly responsible for occasioning the admission of the patient to the hospital for care. When it is chosen in Chapter XXI (Factors influencing health status and contact with health services), the principal diagnosis may be precised by a related diagnosis giving the aetiology. Table 2 gives an example of an inpatient record in the PMSI database.

A national scale of costs per DRG for hospitals is computed each year, by measuring costs of stays in about 50 hospitals. Since 2001, an anonymised patient identifier makes it possible to link all the stays of a given patient in different hospitals.

Building the trajectory of care

We used data from the National PMSI database including all hospitalizations in public or private hospitals for years 2008 to 2010. Data were obtained after approval from the CNIL, the so-called French Data Protection Authority.

Data processing steps are described on Figure 2. For the first step (1), all stays for surgical breast cancer treatment in 2009 were identified. A stay was selected by the co-occurrence of an ICD-10 breast cancer code (C50* or D05*) and a CCAM code of breast surgery. For a given patient, the first identified stay in 2009 was considered as the index stay (setp 2). We looked for previous similar situation in 2008 to check that this index stay was indeed a as far as possible a "new case". Step 3 consisted in building the one year care trajectory for each patient identified

Table 2 A PMSI record

Patient ID	XXXXX
Hospital id	54000278
Stay index	12235
Age	56
Gender	F
Admission month	3
Admission year	2009
Admission status	home
Stay duration	10
Discharge status	home
DRG	09C05V subtotal mastectomies without comorbidities
...	...
Principal diagnosis	C504 : Malignant neoplasm of breast, Upper-outer quadrant of breast
Related diagnosis	NA
Comorbidities	I10 : Essential (primary) hypertension
Procedures	QEFA004 : Lumpectomy

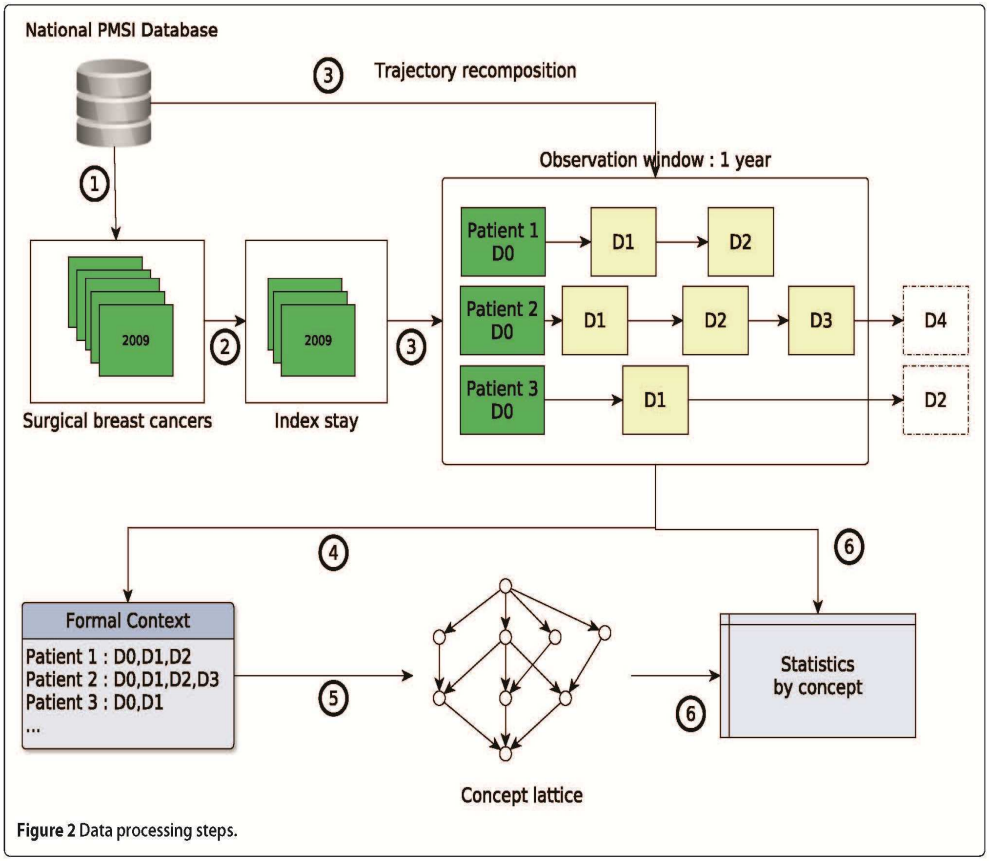


Figure 2 Data processing steps.

at step 2. This trajectory is defined as the sequence of hospitalizations beginning less than 366 days after the index stay, which is the first element of the trajectory. Hence, the observation window covers any stay, for any health condition, occuring within one year from the index stay. However, the PMSI does not apply to ambulatory radiotherapy session in private facilities, though they represent nearly half of the settings in France. We therefore removed all the ambulatory radiotherapy sessions from the analysis.

Several indicators were recorded for each patient: hospitalization costs, number of stays, cumulative length of stay, number of chemotherapy sessions, death. Hospitalization costs were computed using the national scale of costs for public hospitals. Death could only be identified if happening at hospital using the discharge status.

Conceptual clustering of trajectories of care

All the principal and related diagnoses codes in the trajectories were used to build a formal context having patients as objects and diagnoses as attributes (step 4). An excerpt of this context is shown in Table 3. For granularity and tractability reasons, only the first 3 digits of codes were used, except for Z-codes from the chapter XXI of the ICD-10 as 4 and 5 digits convey interesting information (see Additional file 1). A concept lattice was then built using the Coron System (available at http://coron.loria.fr) (step 5). In that lattice, each concept intent can be seen as a condition profile of the patients in the corresponding extent.

The resulting lattice holds all possible combinations of diagnoses that are really observed in the patient trajectories. In order to be interpreted, the lattice is filtered according to concepts support and proportion of own objects. Unfrequent or unstable concepts are removed and the remaining ones are manually reviewed by medical experts.

Table 3 An excerpt of the trajectories formal context

Objects (Patient IDs)	Attribute list (Diagnosis codes)
1	C50, Z768
2	C50
3	C50, R02, Z511
4	C50, I80, Z511
...	...
57552	D05

For each concept, summary statistics are computed considering the patients in its extent: death rates, mean hospitalization costs, mean number of stays, mean cumulative length of stay, mean number of chemotherapy sessions, gender frequencies (step 6).

Though the complete concept lattice can produce non-overlapping classes, filtering the most interesting concepts leads to non-disjoint classes. Actually, a patient appearing in a given concept counts for the computation of statistics in all of its superconcepts. Meanwhile, if necessary, FCA can be combined with other techniques to produce disjoint groups. In this work, we used a regression tree analysis to explain the effects of each profile on the total care cost [32]. In this analysis, each concept is considered as a predictor : its value is set to TRUE if the patient appears in the concept extent. A tree was then grown using the R package rpart [33,34] and was evaluated using 10-fold cross-validation to estimate the Mean Square prediction Error (MSE).

Results

57,552 patients were identified by the selection algorithm. The formal context had 1032 attributes and the resulting lattice had 9,159 concepts. The most frequent and most stable concepts were kept (support $\geq$ 300 and own objects proportion $\geq$ 0.1). Table 4 shows the remaining concepts and their related statistics. The most frequent is the concept with an empty intent. This concept holds all the 57,552 patients meaning that no diagnosis code was common to the entire population of the study. This concept can be taken as a baseline for comparison with other morbidity profiles. Its mean cost is 9,600€, including 3,090€ for chemotherapy sessions. The mean number of stays is 2.0 for a mean cumulative length of stay of 7.3 days. Patients had a mean of 2.9 chemotherapy sessions. They were women in 99% of cases with a mean age of 60.4 years.

53,535 patients had a code of invasive breast cancer (C50) and 5,034 had a code of in situ carcinoma of the breast (D05). The concept (C50, D50) show that, for 1,017 patients, both in situ and invasive neoplasms codes were recorded. The highest cost, 26,139€, was observed for the concept (C50, Z515) coding for invasive neoplasm and palliative care. This concept has also the highest death rate (0.69), number of stays (4.5) and length of stay (43.2 days). The lowest cost, 6,957€, corresponds to the concept of in situ carcinomas of the breast (D05). This concept is associated with the lowest number of stays, length of stay and death rate. However, the concept (C50, D05, Z511) indicates that 371 of these patients had also chemotherapy sessions. Moreover, that group is associated with the highest costs of chemotherapy sessions in Table 4.

Table 4 Statistics by concepts

Intent	Patients	Cost	Stays		Chemotherapy sessions		Death rate	Age
	n	€	n	Cum. length	n	Cost	%	
D05	5034	6957	1.9	5.9	0.6	669	0	57.6
C50, H25	517	9499	3.1	8.2	1.1	1165	1	75.9
Ø	**57552**	**9600**	**2.0**	**7.3**	**2.9**	**3090**	**1**	**60.4**
C50	53535	9902	2.0	7.5	3.1	3318	1	60.6
D05, Z421	482	11471	3.1	11.8	0.3	306	0	50.4
C50, D05	1017	12384	2.8	9.5	3.0	3109	1	56.0
C50, Z421	1339	13484	3.2	12.3	2.0	2055	0	51.2
C50, N61	445	15362	3.5	15.3	3.5	3737	1	60.7
C50, Z452, Z511	13214	15736	2.9	7.8	7.9	8341	1	56.1
C50, Z511	20820	16113	2.7	8.4	8.1	8531	1	55.3
C50, C77	863	16590	3.0	9.6	5.9	6266	1	57.6
C50, C77, Z452, Z511	348	17803	3.5	9.4	7.4	7822	1	57.0
C50, D05, Z511	350	18039	3.1	9.7	8.6	9034	1	53.4
C50, C77, Z511	687	18351	3.1	9.6	7.5	7871	1	55.7
C50, Z421, Z511	332	19946	3.6	12.4	7.9	8288	1	48.5
C50, D61, Z452, Z511	420	22319	4.5	15.5	8.1	8531	3	57.4
C50, D61, Z511	622	22598	4.4	16.4	8.0	8396	3	56.6
C50, C79	372	23052	4.0	28.6	6.3	6587	24	58.9
C50, Z515	365	26139	4.5	43.2	4.2	4371	69	65.5

Statistics are computed on a per patient basis.

Patients in concepts with a code of plastic surgery of breast (Z421) were generally younger: 50.4 years for concept (D05, Z421) and 51.2 for (C50, Z421). At the opposite, senile cataract was observed for older patients: 75.9 years for concept (C50, H25). It can be noticed that the cost for this concept (9,499) is slightly lower than the baseline (9,600) but with a higher number of stays (3.1 vs. 2.0).

Concepts holding patients with advanced malignancies such as secondary locations of lymph nodes (C77), or other and unspecified sites (C79) were associated with higher costs and death rates.

The concept (C50, D61, Z511) is a subconcept of (C50, Z511): all the patients in its extension appear also in the extension of (C50, D61). Comparing the costs of these two concepts reveals that aplastic anaemia (D61) is related with an increase of at least 6400€. The hierarchical structure of the lattice, with super/sub concept relation, allows for such comparisons; for example, N61 (Inflammatory disorders of breast) is associated with increased costs (15,362 vs 9,902) and number of stays (3.5 vs. 2.0) when comparing concepts (C50, N61) with (C50).

The patients appearing in the extent of a concept also appear in the extents of all its superconcepts. Because of this, a filtered lattice can not be seen as hierarchy of disjoint classes. Whenever it may be desirable to achieve a partition of the population, the lattice can be used in conjunction with other techniques. Figure 3 shows the results of a regression tree where the dependent variable is the cost and the predictors are the concepts in which trajectories of care lie (found in Table 4). Each node is labeled by: 1-the number of patients in the node, 2-the average cost for the node. Arrows are labelled by concept membership. By convention, patients having the profile represented by a concept are in the upper branch of a subtree. Figure 4 shows the relative reduction of the prediction error according to the number of splits. The first split explains most of the cost variability. After 5 splits, giving the tree in Figure 3, additional predictors have little effect on the accuracy of the model.

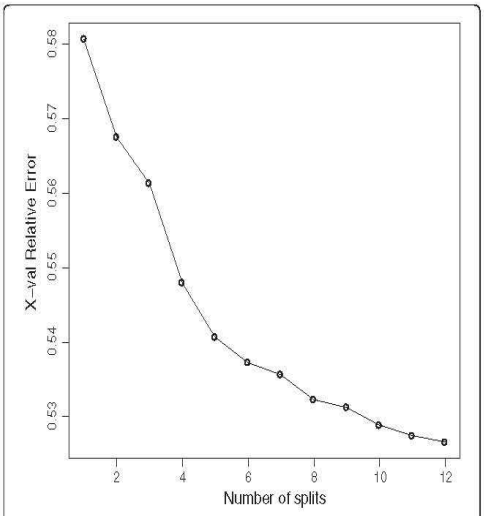


Figure 4 Reduction of the MSE according to the number of splits.

The first split is given by the concept (C50, Z511): invasive BC and chemotherapy session. 20,820 patients shared these two codes with a mean cost of 16,110€. For other patients (n = 36,728), at least one of these two codes had never been used as principal diagnosis. This was associated with a nearly 3 times lower total cost: 5,908€. The other nodes show profiles associated with increased costs: plastic surgery of the breast (Z421), palliative care (Z515) and D61 (aplastic anaemia). Terminal nodes indicate that the majority of the patients (n = 35,562, 62%) are associated with the lowest costs (5,690€). Theses patients either had an in situ carcinoma, or had an invasive cancer but no stay for chemotherapy session. They did not have any stay for plastic surgery nor palliative care. The next most frequent profile is the one of patients with invasive cancer

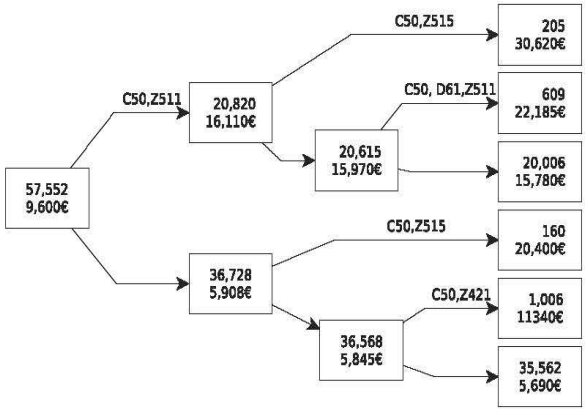


Figure 3 Regression tree.

and chemotherapy sessions (n = 20,006, 35%) for a mean cost of 15,780€.

Discussion

We have presented an approach aiming at clustering care trajectories and analyzing hospitalization costs according to different morbidity profiles. Our method makes use of existing data from the French national casemix system. This has several advantages. First, there is no additional collection of data. Second, this guaranties a form of consistency, quality, homogeneity of data; like other DRG systems, the PMSI is covered by an official guide of data coding, data recording and data transmission rules [35]. Third, this also ensures stability for long term analysis and ulterior comparisons.

Formal Concept Analysis is a powerful conceptual clustering method. Its implementation is fairly simple and requires only a binary table relating individuals and their attributes. It is able to deal with massive amounts of data and is fully unsupervised, thus requiring no labeled training data. In our study, FCA has produced clinically meaningful categories based on sets of diagnoses. These categories are the result of frequent co-occurrences of ICD 10 codes in the patients care trajectories. In this, FCA can be considered as a "white box" model with a straightforward interpretation. We represented diagnoses by their 3-digit ICD codes. A finer grained representation would have resulted in over specific and small categories, sensible to instability of coding. The level of granularity chosen for the attributes is to be taken in consideration to achieve a compromise between generality and specificity of the concepts. As an interesting side effect, FCA can also be used to assess quality of data in the DRG system. We discovered that 1017 patients had both D05 (in situ) and C50 (invasive) breast cancer codes. A closer look at the data revealed that the main diagnosis at inclusion was C50 for 449 (44%) of them, meaning that D05 was recorded subsequently. It is likely that there are coding errors, though we have no evidence at this time to distinguish them from multiple tumors. The concept (C50, D05) could also reflect genuine diagnosis errors. Besides the potential clinical consequences of such errors, our work suggests that they have an economical impact: the cost of concept (C50, D05, Z511) shows a difference of 1,926€ compared to concept (C50, Z511). This kind of fortuitous discovery is actually one of the goals and one of the benefits of data driven approaches.

We analyzed care trajectories in the field of breast cancer. Since 2001, the French casemix information system uses an anonymized identifier linking all the contacts of a given patient with the healthcare system. It is now possible to study groups of patients at a national scale and to analyze resource use in an quasi-exhaustive manner. Thus, our results are not submitted to sampling error. However, they strongly depend on the algorithm used to select patients and build the care trajectory. Several studies have assessed the use of administrative databases to detect cancer cases [12,36,37]. Our method selects patients with surgical breast cancer on a combination of diagnosis and procedure codes which can reduce the false positive rate [38]. The constitution of the care trajectory should be elaborated in order to minimize censoring bias. We standardized the observation window by fixing the start of the trajectory as the first stay for breast cancer surgery and recording all the stays occuring within one year, a sufficient follow-up time to capture the first phase of treatment. The events constituting the care trajectory were any stay recorded by the PMSI, with the exception of ambulatory radiotherapy sessions. This limit in our study is due to the absence of recording of radiotherapy by the PMSI in private settings.

Costs were estimated using the national scale of costs per DRG in hospitals and reflect use of resources by the care providers. Our results show a great variability of costs according to morbidity profiles. Our approach may be used by care providers to take strategic decisions and adapt their resources from a patient-centric point of view, taking into account a whole trajectory rather than a single acute episode of care. Depending on their capacities, they could either evolve towards a more integrated approach, or on the contrary, take advantage of their core competencies and position themselves in the continuum of chronic care. Our method brings also valuable information for hospitals taking part in repetitive treatments (for example annual costs of care per patient, number and total length of episodes of care). Such information is generally unknown from hospital managers because hospital information systems are not aware of what is happening outside the facility in a trajectory of care. At a regional level, this information may help hospital managers to better plan recruitment of patients according to local needs and other hospitals activity. Moreover, as the French system is incited by authorities to develop multidisciplinarity and collaborations, the analysis of care trajectories is essential for implementing more integrated care processes. This work shows how administrative databases and data mining methods can be used to produce descriptions of care trajectories that are both medical and economical. Our approach can highlight discussions and support decisions between partners who are setting up collaborations.

From the institutional standpoint, our method allows health policy makers to set-up cost-of-illness analysis on a national scale. Compared to other patient management systems such as Ambulatory Care Groups [39], it is flexible and reuses existing data, avoiding the drawbacks of a dedicated information system. Though it is not completely disconnected from the DRGs, our grouping of care

trajectories is at first unsupervised and aggregates similar conditions profiles. It can be used in combination with other supervised techniques such as recursive partitioning [32], frequently used in DRG systems [40], to explain costs of care. Analysis of care trajectories could also be used for fair allocation of resources and funding. Indeed, fee-for-service may introduce imbalance funding between the different parts of some trajectories. A global vision of the trajectory can be necessary to redistribute and equilibrate resources between the healthcare providers through prices adjustment, while keeping the overall care expenditure constant.

Our approach is limited by the availability of linkable data. In France, an anonymized identifier can track a patient along its journey through the healthcare system at a national scale. Even though health administrative databases are implemented in many countries, such an identifier may not always exist or may have a limited coverage of the population. Our study focuses on breast cancer but other chronic pathologies can be considered. However, for many of them, care is delivered on an ambulatory basis. In that context, the availability of diagnosis data can be reduced. Eventually, the identification of cases may be subjected to misclassification bias [41,42].

Conclusion

In this paper, we have presented an approach for clustering trajectories of care. Our system is based on Formal Concept Analysis and reuses routinely available claim data. It is flexible and facilitates the longitudinal exploration of treatment practices in the field of chronic diseases. With the example of breast cancer in France, we have demonstrated the possibility of studying trajectories of care at a national level and describing hospitalizations costs according to condition profiles. Classes resulting from an unsupervised process were clinically and economically relevant. Our approach could help healthcare professionals and policy makers to setup cost-of-illness analysis and plan allocation of resources on a patient basis rather than a visit basis.

Additional file

Additional file 1: ICD10 codes used in the article. An CSV file with ICD10 codes and their labels.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NJ formulated the study design, performed data processing and analysis, interpreted the results and drafted the manuscript. GN performed data pre-processing. GN and CQ participated in the study design and interpretation of results. All authors participated in revising the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grants from the Institut National du Cancer (INCa).

Author details

[1]Université de Lorraine, LORIA UMR 7503, F-54000, Nancy, France. [2]CHU de Nancy, Département d'information médicale, F-54000, Nancy, France. [3]CHRU Dijon, Service de Biostatistique et d'Informatique Médicale (DIM), F-21000, Dijon, France. [4]Inserm, U866, Univ de Bourgogne, F-21000, Dijon, France. [5]Université de Bourgogne, F-21000, Dijon, France.

Received: 26 February 2013 Accepted: 20 November 2013
Published: 30 November 2013

References

1. Allotey P, Reidpath DD, Yasin S, Chan CK, de-Graft Aikins A: **Rethinking health-care systems: a focus on chronicity.** *Lancet* 2011, **377**(9764):450–451.
2. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML: **Projections of the cost of cancer care in the united states: 2010-2020.** *J Natl Cancer Inst* 2011, **103**(2):117–128.
3. Gill D, Bruce D, Tan PH: **Controlling the cost of breast cancer.** *Eur J Cancer Care (Engl)* 2011, **20**(6):703–707.
4. Lund JL, Yabroff KR, Ibuka Y, Russell LB, Barnett PG, Lipscomb J, Lawrence WF, Brown ML: **Inventory of data sources for estimating health care costs in the united states.** *Med Care* 2009, **47**(7 Suppl 1):127–142.
5. Yabroff KR, Warren JL, Banthin J, Schrag D, Mariotto A, Lawrence W, Meekins A, Topor M, Brown ML: **Comparison of approaches for estimating prevalence costs of care for cancer patients: what is the impact of data source?** *Med Care* 2009, **47**(7 Suppl 1):64–69.
6. Beckowski MS, Goyal A, Goetzel RZ, Rinehart CL, Darling KJ, Yarborough CM: **Developing alternative methods for determining the incidence, prevalence, and cost burden of coronary heart disease in a corporate population.** *J Occup Environ Med* 2012, **54**(8):1026–1038.
7. Dombkowski KJ, Lamarand K, Dong S, Perng W, Clark SJ: **Using medicaid claims to identify children with asthma.** *J Public Health Manag Pract* 2012, **18**(3):196–203.
8. Bauer HM, Wright G, Chow J: **Evidence of human papillomavirus vaccine effectiveness in reducing genital warts: an analysis of california public family planning administrative claims data, 2007-2010.** *Am J Public Health* 2012, **102**(5):833–835.
9. van Walraven C, Austin PC, Manuel D, Knoll G, Jennings A, Forster AJ: **The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model.** *J Clin Epidemiol* 2010, **63**(12):1332–1341.
10. Aboa-Eboulé C, Mengue D, Benzenine E, Hommel M, Giroud M, Béjot Y, Quantin C: **How accurate is the reporting of stroke in hospital discharge data? a pilot validation study using a population-based stroke registry as control.** *J Neurol* 2013, **260**(2):605–613.
11. Quantin C, Benzenine E, Ferdynus C, Sediki M, Auverlot B, Abrahamowicz M, Morel P, Gouyon JB, Sagot P: **Advantages and limitations of using national administrative data on obstetric blood transfusions to estimate the frequency of obstetric hemorrhages.** *J Public Health (Oxf)* 2013, **35**(1):147–156.
12. Quantin C, Benzenine E, Hägi M, Auverlot B, Abrahamowicz M, Cottenet J, Fournier E, Binquet C, Compain D, Monnet E, Bouvier AM, Danzon A: **Estimation of national colorectal-cancer incidence using claims databases.** *J Cancer Epidemiol* 2012, **2012**:298369.
13. Husain MJ, Brophy S, Macey S, Pinder LM, Atkinson MD, Cooksey R, Phillips CJ, Siebert S: **Herald (health economics using routine anonymised linked data).** *BMC Med Inform Decis Mak* 2012, **12**:24.
14. Fetter R, Shin Y, Freeman J, Averill R, Thompson JD: **Case mix definition by diagnosis-related groups.** *Med Care* 1980, **18**(2):1–53.
15. Fayyad U, Piatetsky-Shapiro G, Smyth P: **The kdd process for extracting useful knowledge from volumes of data.** *Commun ACM* 1996, **29**(11):27–34.
16. Wille R: **Restructuring lattice theory: an approach based on hierarchies of concepts.** In *Ordered Sets. NATO Advanced Study Institutes Series, vol. 83*. Springer Netherlands: Reidel; 1982.
17. Priss U: **Formal concept analysis in information science.** *Ann Rev Information Sci Technol* 2006, **40**:521–543.

18. Agrawal R, Imielski T, Swami A: **Mining association rules between sets of items in large databases.** In *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data.* New York: ACM; 1993:207–216.
19. Pasquier N, Bastide Y, Taouil R, Lakhal L: **Efficient mining of association rules using closed itemset lattices.** *J Info Syst* 1999, **24:**25–46.
20. Zaki MJ, Hsiao CJ: **Charm: an efficient algorithm for closed itemset mining.** In *SDM.* Edited by Grossman RL, Han J, Kumar V, Mannila H, Motwani R. Arlington: SIAM; 2002.
21. Wang J, Han J, Pei J: **Closet+: searching for the best strategies for mining frequent closed itemsets.** In *KDD.* Edited by Getoor L, Senator TE, Domingos P, Faloutsos C: ACM; 2003:236–245.
22. Valtchev P, Missaoui R, Godin R: **Formal concept analysis for knowledge discovery and data mining: the new challenges.** In *ICFCA Lecture Notes in Computer Science, vol. 2961.* Edited by Eklund PW. Berlin, Heidelberg: Springer; 2004:352–371.
23. Cole R, Eklund P: **Scalability in formal concept analysis.** *Comput Intell* 1999, **15:**11–27.
24. Jiang G, Ogasawara K, Endoh A, Sakurai T: **Context-based ontology building support in clinical domains using formal concept analysis.** *Int J Med Inform* 2003, **71**(1):71–81.
25. Jay N, Kohler F, Napoli A: **Using formal concept analysis for mining and interpreting patient flows within a healthcare network.** In *Concept Lattices and Their Applications. Lecture Notes in Computer Science, vol. 4923.* Edited by Yahia S, Nguifo E, Belohlavek R. Berlin, Heidelberg: Springer; 2008:263–268.
26. Aswani Kumar C, Srinivas S: **Mining associations in health care data using formal concept analysis and singular value decomposition.** *J Biol Syst* 2010, **18**(04):787–807.
27. Kaytoue M, Kuznetsov SO, Napoli A, Duplessis S: **Mining gene expression data with pattern structures in formal concept analysis.** *Inf Sci* 2011, **181**(10):1989–2001.
28. Kumar CA: **Fuzzy clustering-based formal concept analysis for association rules mining.** *Appl Art Intell* 2012, **26**(3):274–301.
29. Stumme G, Taouil R, Bastide Y, Pasquier N, Lakhal L: **Computing iceberg concept lattices with titanic.** *Data Knowl Eng* 2002, **42**(2):189–222.
30. Kuznetsov S, Obiedkov S, Roth C: **Reducing the representation complexity of lattice-based taxonomies.** In *Proc. of ICCS 15th Intl Conf Conceptual Structures. LNCS/LNAI vol. 4604.* Edited by Priss U, Polovina S, Hill R. Berlin, Heidelberg: Springer; 2007:241–254.
31. Jay N, Kohler F, Napoli A: **Analysis of social communities with iceberg and stability-based concept lattices.** In *International Conference on Formal Concept Analysis (ICFCA'08). Lecture Notes in Artificial Intelligence, vol. 4933.* Berlin, Heidelberg: Springer; 2008:258–272.
32. Breiman L, Friedman J, Stone CJ, Olshen RA: *Classification and Regression Trees.* New York: Chapman and Hall/CRC; 1984.
33. Core Team R: *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing; 2013. http://www.R-project.org/.
34. Therneau T, Atkinson B, Ripley B: **Rpart: Recursive Partitioning, (2013). R package version 4.1-3.** http://CRAN.R-project.org/package=rpart.
35. Ministère des affaires sociales et de la santé: **Guide méthodologique de production des informations relatives à l'activité médicale et à sa facturation en médecine, chirurgie, obstétrique et odontologie.** Technical Rep, Bulletin officiel, 2012/6 bis, Fascicule spécial 2012.
36. Couris CM, Schott AM, Ecochard R, Morgon E, Colin C: **A literature review to assess the use of claims databases in identifying incident cancer cases.** *Health Serv Outcomes Res Method* 2003, **4**(1):49–63.
37. Mitton N, Colonna M, Trombert B, Olive F, Gomez F, Iwaz J, Polazzi S, Schott-Petelaz AM, Uhry Z, Bossard N, Remontet L: **A suitable approach to estimate cancer incidence in area without cancer registry.** *J Cancer Epidemiol* 2011, **2011:**418968.
38. Quantin C, Benzenine E, Fassa M, Hägi M, Fournier E, Gentil J, Compain D, Monnet E, Arveux P, Danzon A: **Evaluation of the interest of using discharge abstract databases to estimate breast cancer incidence in two french departments.** *Stat J IAOS: J Int Assoc Official Stat* 2012, **28**(1):73–85.
39. Starfield B, Weiner J, Mumford L, Steinwachs D: **Ambulatory care groups: a categorization of diagnoses for research and management.** *Health Serv Res* 1991, **26**(1):53–74.
40. Grubinger T, Kobel C, Pfeiffer KP: **Regression tree construction by bootstrap: Model search for drg-systems applied to austrian health-data.** *BMC Med Info Dec Mak* 2010, **10**(1):9.
41. Smidth M, Sokolowski I, Kærsvang L, Vedsted P: **Developing an algorithm to identify people with chronic obstructive pulmonary disease (copd) using administrative data.** *BMC Med Inform Decis Mak,* **12:**38.
42. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A: **Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data.** *J Clin Epidemiol* 2011, **64**(8):821–829.

doi:10.1186/1472-6947-13-130
Cite this article as: Jay *et al.*: A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. *BMC Medical Informatics and Decision Making* 2013 **13**:130.

VII. COMMUNICATIONS ORALES

- G. Nuemi, F. Afonso, C. Toque, M. Touati, E. Diday, C. Quantin. Symbolic Data Analysis of Cancer Care Trajectories in the region of Burgundy: Application to Lung Cancers. Workshop in Symbolic Data Analysis, Namur, Belgium, 2011.
- G. Nuemi, F. Afonso, A. Roussot, E. Combier, J.-M. Amat-Roze, C. Quantin. Typologie des prises en charge du cancer du poumon chez les patients résidant dans la région Bourgogne. ADELF-EMOIS2012, 2012, Dijon, France. *Revue d'épidémiologie et de santé publique, Vol. 60 N° S1 p. S14*.
- N. Jay, E. Egho, G. Nuemi, F. Kohler, A. Napoli, C. Quantin. Apports des méthodes de fouille de données pour l'étude des trajectoires de prise en charge du cancer du poumon en région Bourgogne. ADELF-EMOIS2012, 2012, Dijon, France. *Revue d'épidémiologie et de santé publique, Vol. 60 N° S1 p. S13-S14*.
- A. Roussot, E. Combier, G. Nuemi, J.-M. Amat-Roze, C. Quantin. Analyse spatiale des trajectoires de prise en charge des patients atteints de cancer primitif du poumon en région Bourgogne. Journées du Journal de Gestion et d'Économie Médicales 2012, Paris, France.
- G. Nuemi, F. Afonso, A. Roussot, A.-M. Bouvier, C. Quantin. Typologie des prises en charge du cancer colorectal chez les patients résidant dans la région Bourgogne. EMOIS, 2013, Nancy, France. Revue d'épidémiologie et de santé publique, Vol. 61 N° S1 p. S9-S10.
- Elias Egho, Nicolas Jay, Chedy Raïssi, Gilles Nuemi, Catherine Quantin, Amedeo Napoli: An Approach for Mining Care Trajectories for Chronic Diseases. 14ème Conférence internationale Artificial Intelligence in Medicine (AIME 2013), Murcia, Espagne, mai 2013.
- Roussot, G. Nuemi, A-M. Bouvier, E. Combier, J.M. Amat-Roze, C. Quantin, Analyse spatiale des prises en charge chirurgicales des patients atteints de cancer colorectal en région Bourgogne, congrès international PCSI (Patient Classification Systems International), à Avignon, 17 – 20 octobre 2012.

VIII. ANNEXES

Annexe 1 : Description d'un épisode de prise en charge et d'une trajectoire de soin

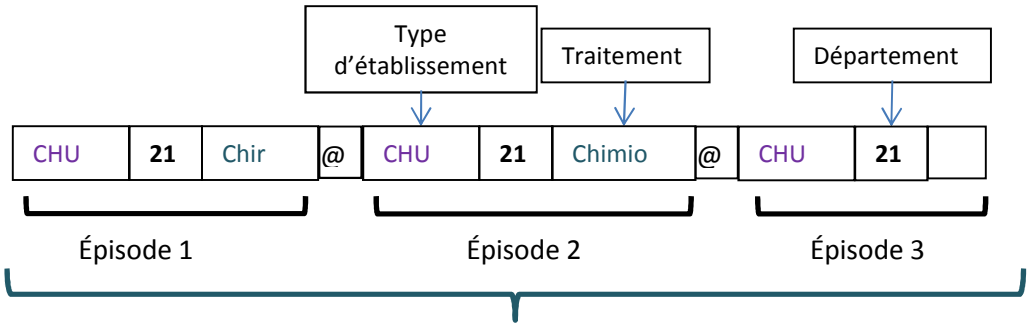
Épisode de prise en charge

C'est une synthèse du séjour hospitalier avec les éléments suivant :

- Le type d'établissement fréquenté : CHU, centre hospitalier, centre de lutte contre le cancer (CLCC), cliniques et les établissements privés participants au service public (PSPH)
- Le département de localisation de cet établissement
- Le type de prise en charge thérapeutique contre le cancer reçu : Chirurgie, chimiothérapie ou radiothérapie.

-

- **Trajectoire de soins**
- Elle représente la séquence chronologique des différents épisodes de prise en charge d'un patient.
- ***Exemple***



Exemple de <u>trajectoire hospitalière</u> pour un patient

-

- **Variables illustratives des différentes typologies de prise en charge**

Nom de la variable[1]	Localisation tumorale	Description
département	- Poumon - Colon-rectum - Sein	Département de résidence lors de la 1ère chirurgie contre le cancer
sexe	- Poumon	
distance	- Poumon - Colon-rectum - Sein	Distance parcourue en voiture et en privilégiant les autoroutes ; Calculée comme la somme sur tous les épisodes de prise en charge (PEC) comme la distance entre le lieu de résidence du patient et la localisation de l'entité juridique de l'établissement de santé où à eu lieu le soin.
traitement	- Poumon - Colon-rectum - Sein	Types de PEC (leur association) liés au cancer (chirurgie chimiothérapie, radiothérapie)
premierEtab	- Poumon - Colon-rectum - Sein	Le type de l'établissement[2] de santé où a eu lieu la 1ère intervention chirurgicale contre le cancer.
âge	- Colon-rectum - Sein	
hôpitaux	- Poumon - Colon-rectum - Sein	Types des différents hôpitaux fréquentés après le premierEtab.
chirurgie	- Poumon - Colon-rectum - Sein	Description de l'acte de chirurgie réalisée lors de la 1ère intervention contre le cancer.
typeTraj	- Poumon - Colon-rectum - Sein	Caractérisation de la trajectoire de soins en termes de : 1. **Fuite régionale** : lorsqu' au moins un épisode de PEC s'est déroulé en dehors de la région Bourgogne 2. **Régionale** : lorsqu'il n'y a pas eu de fuites, mais au moins un épisode de PEC en dehors du département de résidence du patient (enregistré lors de la 1ère intervention chirurgicale) 3. **Proximité ou territoriale** : l'ensemble des épisodes de prises en charge se sont déroulés dans des établissements localisés dans le même département que celui de résidence du patient.
dureeSejour	- Poumon - Colon-rectum	La somme en nombre de jours des durées d'hospitalisation de chacun des épisodes de PEC constituant la trajectoire de soin du patient.
nbepisode	- Colon-rectum - Sein	Nombre d'épisodes de PEC composant la trajectoire du patient

[1] *Voire les figure 1-4-7*

[2] *Les types d'établissements : CHU, centre hospitalier(CH), clinique (CL), Soins de suite (SSR), établissements privés participant au service public (PSPH)*

Annexe 2 : Flowchart décrivant le circuit des données depuis la base nationale

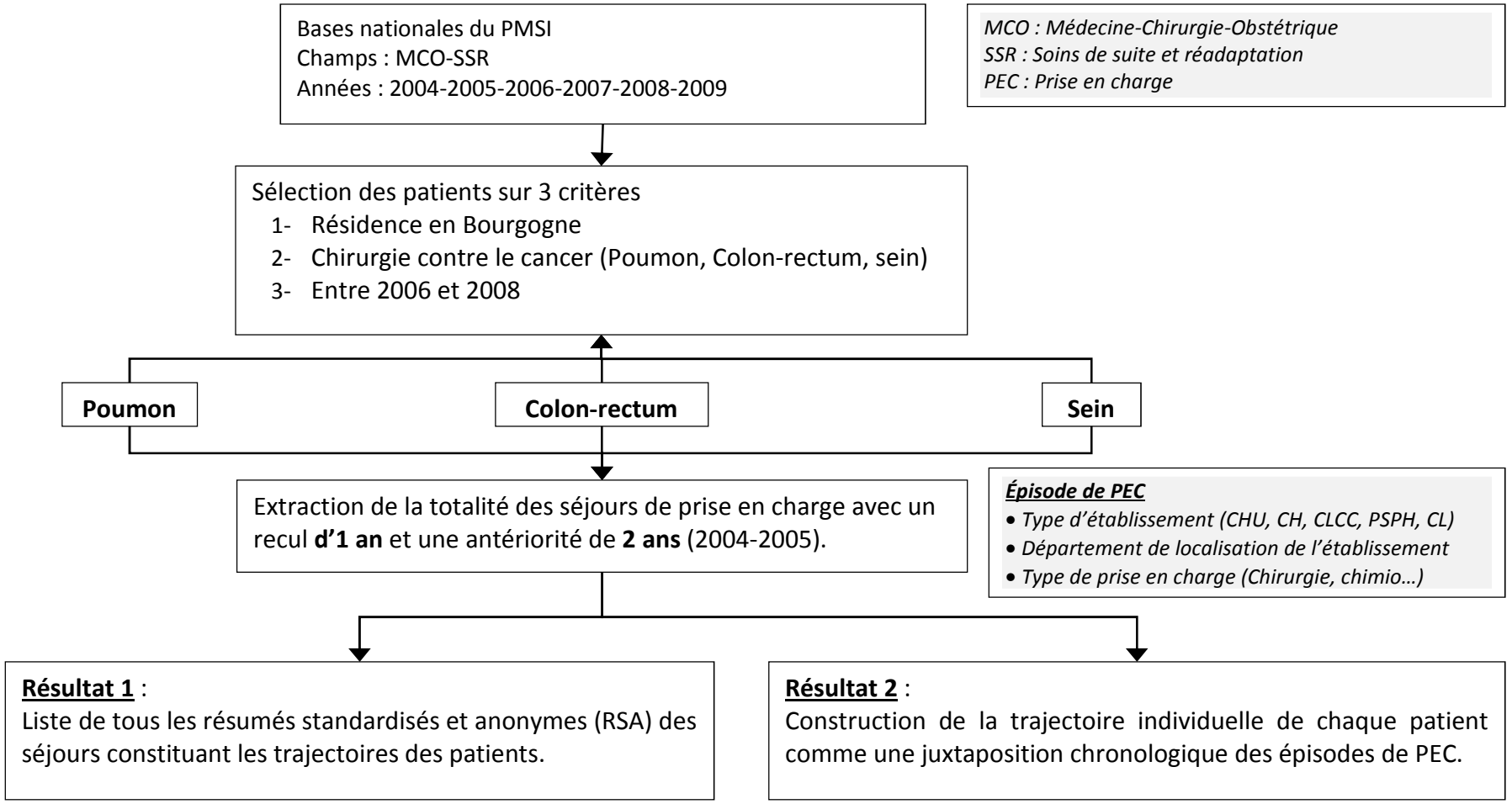


Figure 9 : Flowchart décrivant une première partie du circuit des données d'analyse

Annexe 3 : Flowchart décrivant l'extraction des données patients/séjours pour les cancers du poumon, du colon-rectum et du sein

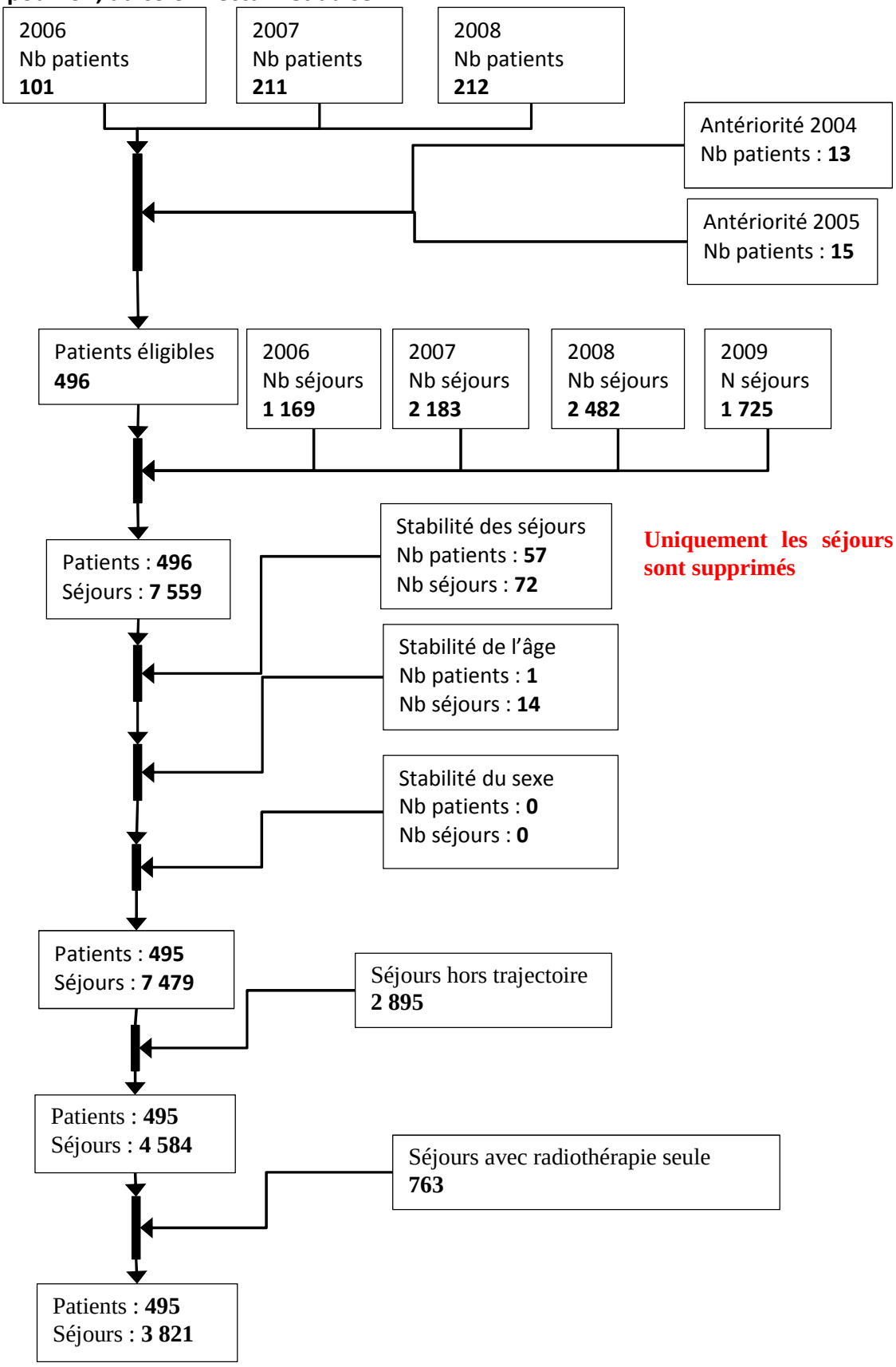


Figure 10 Flowchart décrivant l'extraction des données patients/séjours pour le cancer du poumon.

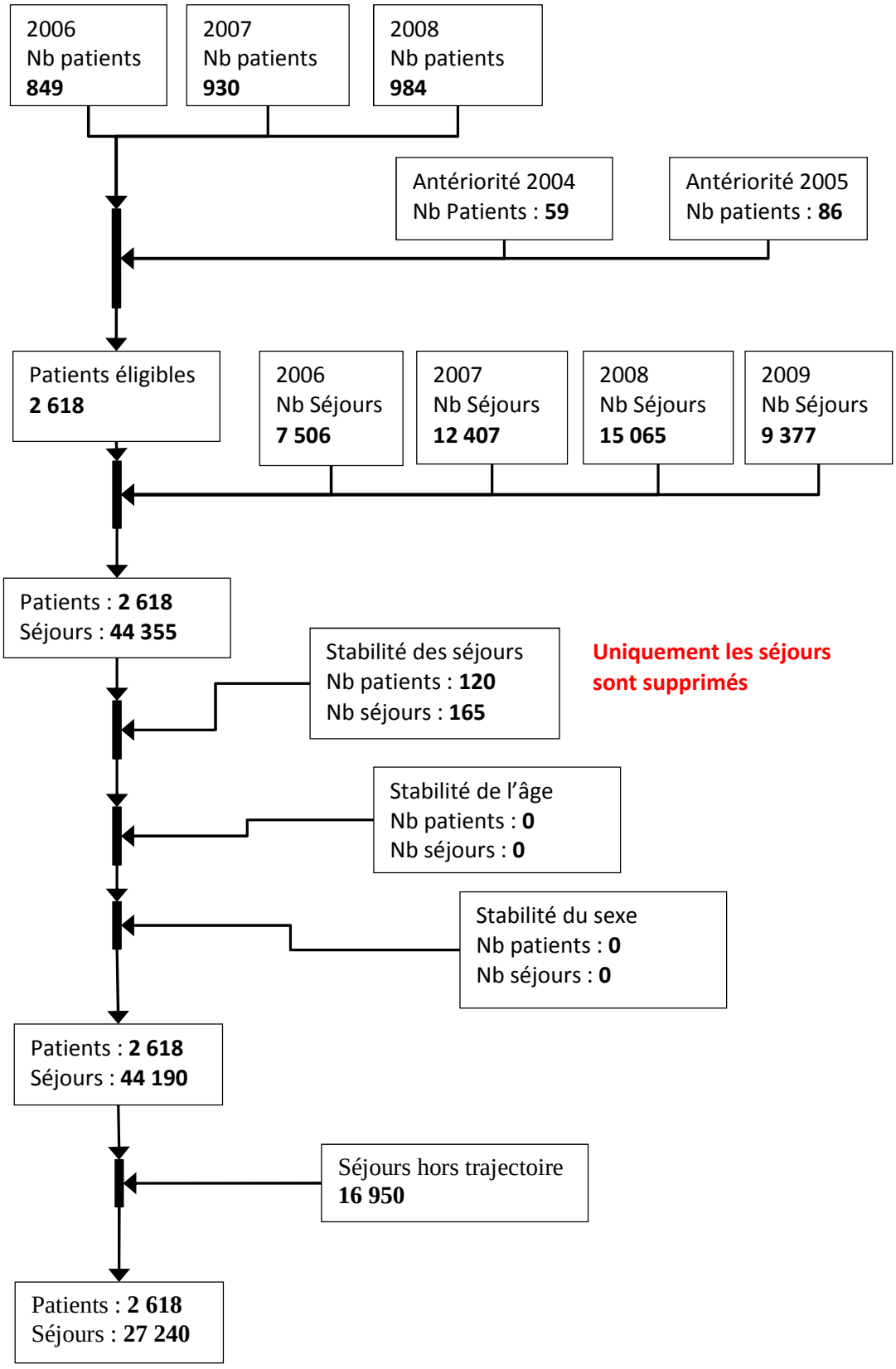


Figure 11: Flowchart décrivant l'extraction des données patients/séjours pour le cancer colorectal.

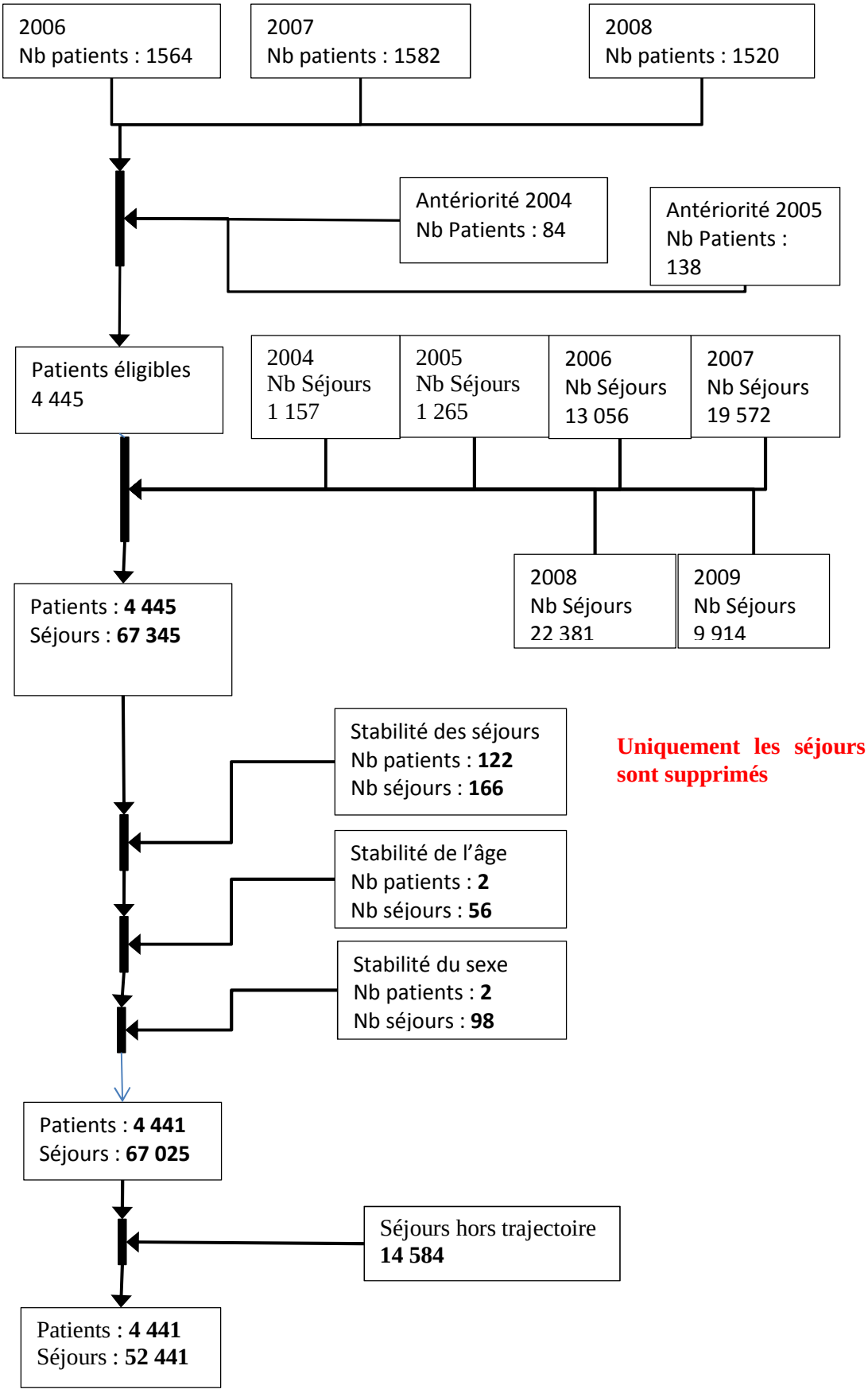


Figure 12: Flowchart décrivant l'extraction des données pour le cancer du sein.

Annexe 4

1. MÉTHODE d'extraction des données séjours/patients et de reconstitution de la trajectoire avec un recul d'un an : Exemple du cancer du sein

   a. Les outils

Dénomination	**Description**
GeriPMSI	Module dédié d'extraction des séjours de la base nationale du PMSI.
SAS®	Pour le traitement des données : les sorties de GeriPMSI entre autres
R®	Analyse préliminaires des séquences de prise en charge avec construction d'une matrice avec les taux de transition entre les différents épisodes de prise en charge
Google® Maps	Extraction des informations de géolocalisation des différents établissements impliquées dans la prise en charge des patients.

   b. Les différents phases du processus

      i. Phase 1

- Sélection des patients vérifiant les critères d'inclusion
- Construction de la liste définitive des patients
- Extraction de la totalité des séjours appartenant à ces patients dans les champs MCO-SSR (HAD mais non utilisées) pour la période concernée.
- Calcul des distances résidence-établissement
- Calcul de l'indice de comorbidité (indice d'Elixhauser)

      ii. Phase 2

Application :

1- des contrôles qualités sur les séjours,

2- Suppression de l'antériorité (2004-2005)

Nb Patients	Nb séjours
4 667	73 471

Nb Patients	Nb séjours
4 441	52 441

      iii. Phase 3

Création de nouvelles variables :

- Distances et temps de parcours depuis leurs lieux de résidence vers l'établissement de prise en charge
- Les classes d'actes de chirurgie
- La trajectoire comme une juxtaposition des différents épisodes de prises en charge

      iv. Phase 4

Cette phase marque le début du pré-traitement de données préalables à la mise en œuvre des outils d'analyse de données symboliques. Il s'agit ici de la création de la variable qui représentera le concept (langage de données symboliques).

CHAPITRE 2 :
CONSTRUCTION DES PROFILS ÉVOLUTIFS DE QUALITÉ DE VIE :
EXEMPLE EN CANCÉROLOGIE DANS UN ESSAI THÉRAPEUTIQUE DE PHASE III

I. INTRODUCTION

Les cancers digestifs sont parmi les cancers les plus fréquents en France (1, 2). Ceux métastatiques ou localement avancés ont un pronostic sombre. Les traitements proposés n'améliorent que marginalement la survie mais, permettent un progrès en terme de qualité de vie (QdV) des patients (3-5). La QdV relative à l'état de santé est actuellement reconnue comme étant un critère essentiel pour l'évaluation des nouveaux traitements (6-8). Dans ce contexte, une description de l'évolution du ressenti des patients en rapport avec une prise en charge thérapeutique lourde devient alors un objectif secondaire systématique dans les essais de phase III. c'est également le cas en routine dans bon nombre d'études pronostiques (9, 10). Cependant, l'hétérogénéité de cette perception individuelle rend souvent très compliquée l'interprétation par les cliniciens, des résultats objectifs de l'étude. Pourtant, la prise en compte de cet élément subjectif devient d'autant plus importante qu'il s'agit d'un essai thérapeutique de phase III réalisé avec des patients porteurs d'un cancer digestive métastatique et dont l'objectif d'amélioration de la QdV est essentiel. Une méthode pour l'évaluation de la QdV consiste à utiliser des outils psychométriques permettant d'appréhender les symptômes ainsi que le bien-être des patients en leur donnant la parole. Il est ainsi possible de collecter des mesures standardisées de QdV relatifs à la santé des patients (11). L'analyse des données de QdV est un processus très complexe. Nous pouvons souligner tout d'abord le caractère subjectif de l'information recueillie (11). Ensuite, son recueil est réalisé à l'aide d'échelles de mesures multidimensionnelles. Et enfin, la nature longitudinales des données rend quasi inéluctable la présence de données manquantes ; les raisons pouvant être diverses et variées (12). L'impact de ces DM du fait de leurs proportions variables doivent pouvoir être prise en compte pour atténuer le risque non négligeable de biais (12). Malgré cette complexité, l'évolution de la QdV des patients dans un contexte défini (essai clinique) doit pouvoir être décrite, présentée et interprétée d'une manière qui soit simple et compréhensible. Les méthodes d'analyses des données longitudinales n'ont cessé d'évoluer depuis une vingtaine d'années et notamment, depuis les méthodes descriptives à partir de mesures de tendances centrales (moyenne, médiane,…) qui occultaient la nature longitudinale des données puis, l'apport des modèles hiérarchiques qui ont permis une description plus fine des données mais avec une interprétation des résultats qui restait compliquée pour les non-initiés (13). Aujourd'hui, des méthodes encore plus sophistiquées permettent d'identifier des sous-groupes de population homogènes à partir de variables quantitatives (14-16). Elles sont de plus en plus accessibles et permettent d'obtenir des résultats facilement interprétables. De

façon concomitante, les méthodes de prise en compte (au moment de l'analyse) des données manquantes dans les études cliniques sont en cours d'uniformisation (13, 17-21).

L'objectif de ce travail était de construire et de décrire une typologie des profils évolutifs de QdV après un essai thérapeutique comparatif de phase III chez les patients souffrant d'un cancer digestif métastatique.

II. MATÉRIELS ET MÉTHODES

A. DESIGN DE L'ÉTUDE

Nous avons travaillé avec des données issues de l'essai clinique de phase III (FFCD-03-07) réalisé entre juin 2005 et mai 2010. Il s'agissait d'un essai randomisé multicentrique, ouvert, prospectif, qui comparait l'efficacité de deux séquences de polychimiothérapie (FOLFIRI versus Épirubicine-Cisplatine-Capécitabine (ECX) et vis-versa) chez les patients souffrant d'un adénocarcinome de l'estomac ou du cardia localement avancé ou métastatique. Ces derniers avaient été randomisés en 2 groupes (1:1) pour recevoir soit l'ECX comme traitement de première ligne suivie par FOLFIRI en deuxième ligne (bras A), ou la séquence inverse (FOLFIRI en première ligne suivie par ECX comme deuxième ligne) (bras B), selon les critères de stratification suivantes: l'indice de performance de l'OMS (0-1/2); lésions mesurables ou évaluables; centre de recrutement des patients; l'histoire de la chimiothérapie adjuvante ou à la radio-chimiothérapie; localisation de la tumeur (gastrique / cardia); et le type pathologique (linite ou non). Le schéma ECX était composé d'Épirubicine 50 mg/m2 (en 15 min par voie intraveineuse (IV) en perfusion) plus Cisplatine 60 mg/m2 (comme une perfusion IV de 1 h) le Jour 1 suivi par voie orale Capécitabine 1 g/m2 deux fois/jour à partir de J2 jusqu'à J15 toutes les 3 semaines; la dose d'Épirubicine cumulée maximum autorisée était de 900 mg/m2. Le schéma FOLFIRI était composé de l'Irinotécan à 180 mg/m2 (en 90 min perfusion IV) et l'acide folinique 400 mg/m2 (comme une perfusion intraveineuse de 2 h), suivi d'une injection en IV de 400 mg/m2 de 5-FU en bolus puis d'une perfusion sur 46h de 2400 mg/m2 de 5-FU toutes les 2 semaines. Une modification de posologie, hydratation appropriée, et une prémédication avait été prédéfinies dans le protocole de l'étude. Le traitement de première ligne était administré jusqu'à la progression de la maladie, une toxicité inacceptable, le retrait du consentement, ou la mort du patient. Le traitement de deuxième ligne était alors administré après le traitement de première ligne après un minimum de 3 semaines d'intervalle sans traitement, et une récupération clinique et biologiques. Cet essai avait montré une diminution significative du temps jusqu'à échec thérapeutique en première ligne (TET_L1) en faveur du FOLFIRI en première ligne.

B. RECUEIL DES DONNÉES DE QUALITÉ DE VIE

Un objectif secondaire de l'étude clinique était d'évaluer la qualité de vie mesurée par l'auto-questionnaire EORTC-QLQ-C30 (European Organisation for Research and Treatment of

Cancer- Quality of life questionnaire core 3.0). Le déroulement de l'étude prévoyait une évaluation toutes les 8 semaines de l'auto-questionnaire de QdV.

C. QUESTIONNAIRE EORTC QLQ-C30

Le QLQ-C30 est une échelle de mesure qui comporte 30 items ou questions pertinents permettant de mesurer l'état de santé perçus par un large éventail de patients atteints de cancer. Parmi ces 30 items, 17 sont regroupés en 5 dimensions fonctionnelles (physique, cognitif, sociale, émotionnel et vie quotidienne) et un état de santé global/échelle de qualité de vie. Les 13 autres éléments sont regroupés en des dimensions de mesure des symptômes liés au cancer (fatigue, nausées/vomissements, douleurs, dyspnée, diarrhée, insomnie, perte de l'appétit, et constipation). Le questionnaire est validé chez les patients atteints de cancers de l'estomac (22). Pour chaque dimension, un score est calculé selon une méthode standardisée (23). Les scores pour les dimensions fonctionnelles. Par contre, pour les dimensions des symptômes, les scores varient de 0 (aucun retentissement) à 100 (retentissement maximum). Nous avons étudié la dimension fonctionnelle physique qui est basée sur 5 questions (la n°1 à la n°5). Pour les patients, cela signifie évaluer par rapport à la semaine écoulée leur capacité à effectuer certaines activités quotidiennes de la vie comme s'habiller, l'hygiène personnelle, ou faire une promenade.

D. DES PROFILS INDIVIDUELS DE SCORES AUX PROFILS ÉVOLUTIFS DE QUALITÉ DE VIE

Le protocole de l'étude prévoyait après 14 mois de suivis, au moins 7 auto-questionnaires remplis toutes les 8 semaines depuis la date de randomisation par chaque patient. Les 7 premières mesures du scores correspondant à la dimension physique (QLQ-C30 PF2) et notées T1, T2, …, T7 ont été analysées. Un profil individuel de score (PIS) était défini comme la suite chronologique des différents scores calculés à chaque temps pour un patient donné. Partant des PIS, quatre étapes ont été nécessaire jusqu'à la construction des profils évolutifs (PE) de qualité de vie. Celles-ci sont résumées à l'annexe 1 :

1/ La première était un processus d'augmentation des données (les scores) .(18, 24). En effet, l'absence de réponse à au moins 3 questions dans la dimension physique du QLQ-C30 ou l'absence du questionnaire complet se traduisait par un score manquant dans le PIS du patient. Nous avons observé 3 modes de distribution des scores manquants dans un PIS donnés : une distribution dite monotone, c'est-à-dire qu'après un premier score manquant, aucun autre score n'était observée dans le PIS du patient. Ensuite, une distribution intermittente où un

score pouvait être précédé et/ou suivi par une valeur manquante. Et enfin une distribution mixte qui est un mélange des 2 cas précédents (voir annexe 2). Nous avons pris en compte la présence de ces scores manquants en appliquant la méthode de l'imputation multiple (IM) sur les scores. Cette méthode est habituellement recommandée pour les taux importants de données manquantes (13, 17-21). Cette stratégie avait comme intérêt majeur de nous permettre de mettre en œuvre les traitements statistiques envisagés. Ainsi, 100 jeux de données avec des PIS complets ont été générés. À noter que nous n'avons pas réalisé d'imputation après le décès et que les scores ont été alors fixés à une valeur constante.

2/ Dans la deuxième étape, des paramètres de variabilité étaient calculés sur chacun des 100 jeux de données. Ces paramètres ont été proposées par l'équipe Leffondré et al en 2004 pour l'identification des profils de changement de scores de qualité de vie recueillis de manière longitudinale (14). Il s'agissait de 27 paramètres statistiques simples répartis en 3 groupes : d'abord les paramètres décrivant la linéarité du PIS (exemple : l'écart-type, la pente d'une droite ou encore la part de variance expliquée par un modèle linéaire), ensuite ceux reflétant la non-linéarité des PIS comme des changements abrupt sur de courtes périodes (exemple : la moyennes des différences successives entre 2 scores consécutifs, l'écart-type correspondant, la moyenne des différences successives entre 2 scores pris une fois sur 2, etc). Et enfin les paramètres mesurant le contraste entre 2 périodes définies dans un PIS (rapport entre le changement avant sur le changement après). L'annexe 3 présente un extrait de ces paramètres. Ainsi, chaque PIS était alors décrit par ces paramètres de variabilités calculés ; les 100 jeux de données précédents devenant 100 jeux différents des paramètres calculés pour chaque PIS.

3/ La troisième étape consistait à construire des groupes homogènes de PIS. La méthode de classification non supervisée, « k-means », avec les distances Euclidiennes a été appliquée à chaque jeu de données (20, 25). Le choix à priori du nombre de classe ainsi que la sélection des paramètres les plus pertinent pour une classification optimale ont été réalisées automatiquement en suivant une méthodologie standardisée (20, 25, 26). Nous avons ainsi maximisé le critère de classification : le CritCF (26) pour chaque processus de classification appliqué à un jeu de données. Après cette 3° étape, nous disposions de 100 résultats de classification. Chacune ayant été réalisée pour un même nombre de classe et sur les mêmes paramètres de variabilités.

4/ La dernière étape consistait à agréger les 100 résultats de classifications issus de l'étape précédentes en une seule classification. Cette dernière représentant alors ce que nous avions appelé les profils évolutifs (PE) de qualité de vie. Le processus d'agrégation était réalisée de

la manière suivante : L'affectation définitive d'un patient à un PE était donnée par son affectation majoritaire sur les 100 classifications (20). Pour la description de chaque PE à partir des PIS des patients, nous avons utilisé les moyennes des scores pour chaque temps de mesure.

E. ANALYSE DE SENSIBILITÉ

Nous avons réalisé une analyse de la sensibilité de PE obtenu. Il s'agissait d'appréhender l'impact de la distribution des scores manquants imputés. Pour cela, nous avons tout d'abord calculé une probabilité d'affectation d'un patient à une classe spécifique au regard de l'ensemble de ses 100 classifications différentes. Ensuite, nous avons pour chaque PE, représenté graphiquement en utilisant les « boîtes à moustaches » la distribution des probabilités de chacune des classes d'affectations des patients. Et enfin, nous avons utilisé une variante du coefficient de concordance, l'indice ajusté de Rand pour analyser l'affectation finale des patients par rapport aux affectations issues des 100 jeux de données. Nous avons calculé une moyenne du coefficient et sont intervalle de confiance à 95% (27-29)

F. DESCRIPTION DES PROFILS ÉVOLUTIFS

Les variables utilisées pour la description des différents profils étaient réparties en 3 groupes. D'abord les variables individuelles des patients (âge et sexe), ensuite celles correspondant aux informations recueillies à la randomisation : la localisation de la tumeur (Estomac ou Cardia), son type (Linite ou non), le bras de randomisation (FOLFIRI/ECX ou ECX/FOLFIRI) et l'indice de performance de l'OMS (0-1 ou 2). Et enfin, les variables relatives au suivi dans l'étude : le TET_L1, la déclaration d'effet indésirable grave (Oui/Non), la réalisation de la 2ième ligne thérapeutique (Switch Oui/Non) et l'état du patient à la date des dernières nouvelles (Vivant ou décédé). Pour chaque profil évolutif, ces variables étaient représentées sous forme d'histogramme de fréquences où les barres correspondaient à une modalité et la hauteur était la fréquence relative de cette modalité dans les différents profils évolutifs (30).

Les comparaisons de fréquences étaient réalisées avec le test du Chi-2 ou le test exact de Fisher le cas échéant. Les taux de survies étaient estimés avec la méthode de Kaplan-Meier. Pour l'ensemble des tests statistiques réalisés, Le seuil de significativité statistique était fixé à 0,05.

Les analyses statistiques ont été réalisées avec le logiciel de statistique SAS version 9.3. Les représentations graphiques sous forme de barres d'histogrammes étaient réalisées à partir d'outils dédiés.

III. RÉSULTATS

L'essai clinique avait inclus 416 patients randomisés de manière équilibrée (environ 50% par bras). Le bras d'affectation initial définissait pour chaque patient la première ligne thérapeutique. Les patients du bras A recevait l'association ECX et ceux du bras B recevait le FOLFIRI. Les hommes étaient majoritairement représentés quel que soit le bras (environ 74%). Le taux de décès après 14 mois de suivi était estimé à 55 % quels que soit le bras. Les TET_L1 observés étaient plus élevé dans le bras B, 25 semaines contre 18. Le tableau 1 présente le détail pour les autres caractéristiques des patients.

Le nombre total d'auto-questionnaires exploitables s'élevait à 1023 qui provenaient de 364 patients. On comptait pour chaque patient en moyenne 3±2 questionnaires. Ce nombre variait de 1 à 12. Sur les 7 premières évaluations, 2 912 auto-questionnaires étaient attendus pour l'ensemble des patients ayant reçu au moins une dose d'un traitement (soit 416 patients). Nous avons pu reconstituer au total 1 650 scores (57%). Le taux de scores qualifié de manquant s'élevait à environ 43% (1 262) sachant que l'absence de score après le décès n'était pas considérée comme une donnée manquante (DM). La majorité des DM (55%) avait une distribution intermittente et 34% était mixte.

En ce qui concerne la détermination des valeurs optimales pour le nombre de classe à définir à priori et le nombre de paramètres pour les classifications, nous avons obtenu des résultats. En maximisant le critère de classification CritCF, la répartition des résultats de classification selon un nombre de classe défini à priori était de 6%, 5%, 38% et 51% respectivement pour des regroupements en 2, 3, 4 et 5 classes. C'est-à-dire que nous avions obtenu par exemple pour 6 jeux de données sur les 100 un CritCF maximum avec un regroupement en 2 classes. Au final, avec un CritCF moyen de 0,75 s'étendant de 0,61 à 0,87 nous avons retenu la valeur 4 pour le nombre de classe à définir à priori. Pour déterminer le nombre de paramètres, nous avons donc analysé la distribution des 27 paramètres uniquement sur les jeux de données maximisant le CritCF en 4 classes (voir la figure 1). Nous avons retenu les 13 paramètres les plus fréquentes. Le détail de la liste avec les formules correspondantes est fourni dans l'annexe 3.

Nous avons donc au final construit 4 profils évolutifs. Les 416 patients ont été affectés de façon définitive (après l'étape n°4) dans l'un des 4 PE noté dans la suite P1, P2, P3 et P4 dans les proportions respectives suivantes : 6%, 41%, 19% et 34%. En considérant l'écart maximum (EM) entre les différents scores pour un PE donné, nous avons décrit une typologie

en 3 profil distincts : un profil d'amélioration P1 avec un EM=+25 points (pts), un profil de stabilité P2 (EM=-12pts) et 2 profils de dégradations P3 (EM=-21pts) et P4 (EM=-27pts). La figure 2 montre les courbes d'évolution des profils moyens sur les 7 temps d'évaluations. Chaque courbe est encadrée par 2 autres représentant l'écart-type de la moyenne des scores par profil entre les 100 jeux de données. Le tableau 2 montre une synthèse de ces profils réalisée à partir de certaines variables recueillies (âge, sexe, bras de randomisation, indice de performance de l'OMS, taux de décès observé et taux d'effets indésirables graves). De plus, dans ce tableau, la tendance linéaire de chaque profil est rappelée à l'entête. Une autre description des PE plus visuelle est présentée dans la figure 3. En résumé, la variable TET_L1 permettait de distinguer les 4 PE. Nous pouvions également traduire la distinction apparentes entre les 4 PE de la manière suivante : le profil P1 était majoritairement composé de patients du bras ECX/FOLFIRI, P2 contenait les patients relativement jeunes (âgé entre 44 et 66 ans) avec les TET_L1 les plus élevés (au-delà de 31 semaines). Pour le P3, d'une part une majorité de patients déclarait des effets indésirables graves et d'autres part, les TET_L1 étaient les plus faibles (moins de 11 semaines). Enfin, le P4 rassemblait essentiellement les patients ayant un TET_L1 entre 11 et 31 semaines mais avait également le nombre de décès parmi les plus élevés.

L'impact de la distribution des valeurs imputées est représenté sur la figure 4. Cette dernière montre pour chaque profil évolutif la distribution des classes d'affectation initiale des patients constituant chaque PE. Ainsi, P2 était composé en majorité des patients issus de la classe 2 (à 73% en moyenne), mais également des patients de la classe 1 à 15% et très rarement des patients de la classe 3. De même le coefficient de concordance entre chacune des 100 classification et l'affectation finale était en moyenne de 0,62 ($IC_{95\%}$[0,61-0,63]).

IV. DISCUSSION

Ce travail nous a permis de mettre en évidence 3 résultats intéressant : Nous avons tout d'abord obtenu par l'application d'une méthode de simulation adaptée, l'imputation multiple, la reconstitution d'une base de données exhaustive et exploitable sur les informations de qualité de vie relative à la santé de ces patients. Ceci a permis de juguler le problème épineux des données manquantes quasi-indissociable pour le type de pathologie (cancer métastatique) et de prise en charge. Ensuite, nous avons montré qu'il était possible de construire des profils pertinents d'évolutions de la qualité de vie des patients à partir de mesures répétées de qualité de vie. Les résultats peuvent à postériori paraitre intuitifs et évidents. Ainsi, 3 types de profils ont été décrits : le profil d'amélioration (P1) qui avait le nombre de patient le plus faible et une différence de scores d'environ +25 points entre les extrêmes. Le profil de stabilité (P2) qui regroupait les patients avec un score initial autour de 80/100 et un écart maximal de -12 points. Il s'agissait du profil le plus peuplé (41%). Enfin le profil de dégradation qui était organisé sur 2 groupes se distinguant au niveau du score initial des patients: Dans le premier groupe (P3), le moins peuplé (19%) ce score initial était voisin de 43/100 tandis que dans le second (P4), il était voisin de 82/100. Et enfin, nous avons proposé une représentation graphique pour la description des différents profils à partir des variables individuelles et cliniques recueillies. Pour les variables recueillies à l'inclusion telles que la localisation de la tumeur, son type, l'indice de performance de l'OMS et le sexe, on notait une structure des barres d'histogrammes quasi identique d'un profil à l'autre ; pour la variable, âge, la structure était nuancée avec une prédominance des patients les plus âgés dans les profils P1 et P3. Et, pour les patients relativement jeunes, une majorité dans les profils P2 et P3. Pour les variables recueillies tout au long de l'étude comme la déclaration d'un effet indésirable grave (complications) ou encore le temps jusqu'à échec thérapeutique de la 1ère ligne thérapeutique (TET_L1) les structures des barres d'histogrammes étaient différentes d'un profil à l'autre.

Il nous semble important de rappeler que l'analyse et l'appropriation de cette typologie des profils-ne devraient pas occulter le fait de la gestion sous-jacente des scores manquants ainsi que la méthode utilisée. En effet, le principal atout dans l'utilisation de la méthode de l'imputation multiple réside non pas dans le remplacement des valeurs manquantes mais dans la conservation de la distribution des données en l'occurrence ici dans notre cas, des scores de qualité de vie. Il en résulte un modèle de données permettant une interprétation de la réalité plus vraisemblable car moins biaisé que celui que l'on aurait obtenu en appliquant les méthodes d'imputations dites « classiques » : par la moyenne, la dernière valeur disponible,

etc. Cependant, le modèle, malgré les évolutions conserve un certains nombres d'écueils inhérents à la construction des modèles statistiques. L'autre intérêt de ce modèle est la possibilité de pouvoir réaliser les analyses statistiques prévues. La méthode de construction des profils que nous avons utilisée présente comme atout de produire des profils qui sont à priori indépendants des autres variables cliniques ou individuelles recueillies chez les patients concernés. Ceci ouvre la possibilité d'étude des associations entre ces profils et des variables de l'étude (31) ou encore l'identification des facteurs prédictifs de ces profils. Certaines étapes dans la mise en œuvre de cette méthode sont automatisables et optimisées comme par exemple le choix du nombre de profil. Cependant, pour ce qui est de la sélection des paramètres, l'exercice pourrait dans certains cas se révéler long et fastidieux. Le fait que les profils obtenus ne soient pas à priori liés (par construction) aux autres variables cliniques recueillies nous autorise en quelque sorte à utiliser ces dernières pour une description des différents profils. Il est évident qu'une présentation graphique facilite l'exercice d'interprétation du clinicien et que celle que nous avons proposée (les barres d'histogrammes) permet une visualisation de la variabilité entre les différents profils. Cette variabilité pourrait s'avérer dans certaines situations très difficile à appréhender lorsqu'elle est réduite à la seule valeur des écart-types ou encore des intervalles de confiance. Il s'agit d'un mode de présentation des données qui n'est pas encore répandu dans la littérature médicale(30).

La description des différents profils avec les variables cliniques révélait certaines particularités : Dans le profil P1 (amélioration) par rapport aux autres, les patients du bras A ECX/FOLFIRI étaient nettement majoritaires avec des TET_L1 prédominant entre 11 et 31 semaines. Il s'agissait essentiellement de patients qui n'avaient pas changées de bras d'études durant tout l'essai clinique. Autre particularité, la plus grande proportion des patients avec des TET_L1 supérieurs à 31 semaines était présente dans le profil P2 (stabilité). Cependant, il était très difficile d'associer ce fait avec un bras particulier de traitement. Ce constat était identique chez les patients avec les TET_L1 les plus faibles que l'on retrouvait en majorité dans le profil P3 (dégradation). Ces derniers présentaient de forte proportion pour la déclaration des effets indésirables graves. Le lien entre la perception qu'ont les patients de leur QdV vis-à-vis du traitement anticancéreux reçu ne nous apparaissait pas évident devant ces interprétations. Mais, au regard de la littérature, on pouvait remarquer certains éléments de cohérences : par exemple, les patients qui exprimaient une dégradation de leurs qualité de vie (P3) présentaient également sur le plan clinique des TET_L1 faibles (<11 semaines) et avaient été principalement traités avec une chimiothérapie contenant de l'Épirubicine (bras A,

ECX/FOLFIRI) (5). Autres exemple avec le profil P2 où les « bons » résultats cliniques étaient observés majoritairement chez les patients traités à base d'Irinotécan (bras B, FOLFIRI/ECX) (4).

V. CONCLUSION

Il s'agissait avec ce travail de montrer qu'une autre approche était possible dans l'étude de l'association entre la perception qu'ont les patients de l'évolution de leur qualité de vie vis-à-vis des traitements anticancéreux reçus. Nous nous sommes placé dans une situation défavorable, celle de travailler sur un essai thérapeutique complexe sur un cancer grave et métastatique donc avancé avec un risque élevé et prévisible de données manquantes sur les informations de qualité de vie. Avec une méthodologie rigoureuse, les résultats suggèrent fortement d'abord l'intérêt d'approfondir les analyses (notamment dans l'étude de la sensibilité des résultats), ensuite d'étendre les analyses à d'autres dimensions de l'échelle et enfin de comparer les résultats sur d'autres données d'essai thérapeutiques.

VI. TABLEAUX

Tableau 11: Caractéristiques des patients inclus dans l'étude

Variables	Bras A ECX[1]/FOLFIRI ***n=209***	Bras B FOLFIRI/ECX ***n=207***
Sexe		
Hommes n(%)	154 (74)	155 (75)
Indice de performance de l'OMS à J_0		
=0-1 n(%)	175(83,7)	178(86,0)
=2 n(%)	34(16,3)	29(14,0)
Type de la tumeur		
Linite n(%)	46(22,0)	51(24,6)
Âge (ans)		
m±e-t[2] [min-max]	61±11 [28-84]	61±11 [29-81]
Durée de suivi (mois)		
m±e-t[2] [min-max]	11±10 [0-52]	11±8 [0-42]
EIG[3] n(%)		
Après 7 temps d'évaluation	122 (58)	105 (51)
TET_L1[4] (semaines)		
m±e-t (max)	18±14 (64)	25±23 (137)
Décès n(%)		
Global[5]	175 (84)	180 (87)
Après 7 temps d'évaluation	116 (56)	113 (55)

[1] *ECX : Épirubicine-Cisplatine-Capécitabine*

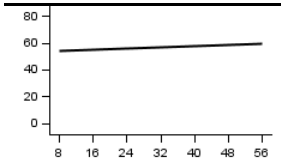
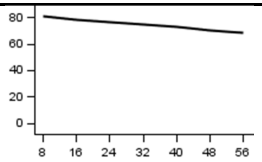
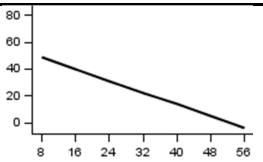
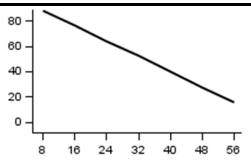
[2] *m±et : moyenne ± écart-type*

[3] *Effet indésirable grave*

[4] *Temps jusqu'à l'échec thérapeutique de la première ligne de traitement*

[5] *Proportion des décès quelle que soit la ligne thérapeutique*

Tableau 12: Caractéristiques des patients par profil évolutif

	Profil 1	Profil 2	Profil 3	Profil 4	
					P[1]
Effectifs	24	171	78	143	
Sexe Hommes n(%)	15 (63)	128 (75)	58 (74)	108 (76)	0,596
Âge (ans) md±inq[2]	63±17	61±15	65±19	60±17	0,370
Indice de performance de l'OMS à J_0 =0-1 n(%) =2 n(%)	 20(83,3) 4(16,7)	 151(88,3) 20(11,7)	 56(71,8) 22(28,2)	 126(88,1) 17(11,9)	 0,004
Type de la tumeur Linite n(%)	7(29,2)	33(19,3)	22(28,2)	35(24,5)	0,369
Bras de randomisation FOLFIRI/ECX n(%)	5(20,8)	93(54,4)	35(44,9)	74(51,8)	0,015
Score Qualité de vie $Score_7$-$Score_1$ (p^*)	21,8(0,520)	13,5(0,012)	-62,7(0,038)	-67,6($<10^{-3}$)	
EIG[3] n(%)	13 (54)	67 (39)	62 (81)	86 (60)	$<10^{-3}$
TET_L1[4] (semaines) md±inq	15±18	30±24	4±7	20±12	$<10^{-3}$
Décès n(%)	3 (13)	18 (11)	78 (100)	131 (92)	$<10^{-3}$

[1] *: p de significativité*

[2] *:médiane± écart inter-quartile*

[3] *Effet indésirable grave*

[4] *Temps jusqu'à l'échec thérapeutique de la première ligne de traitement*

* *: p de significativité de la pente prédite par un modèle linéaire*

VII. FIGURES

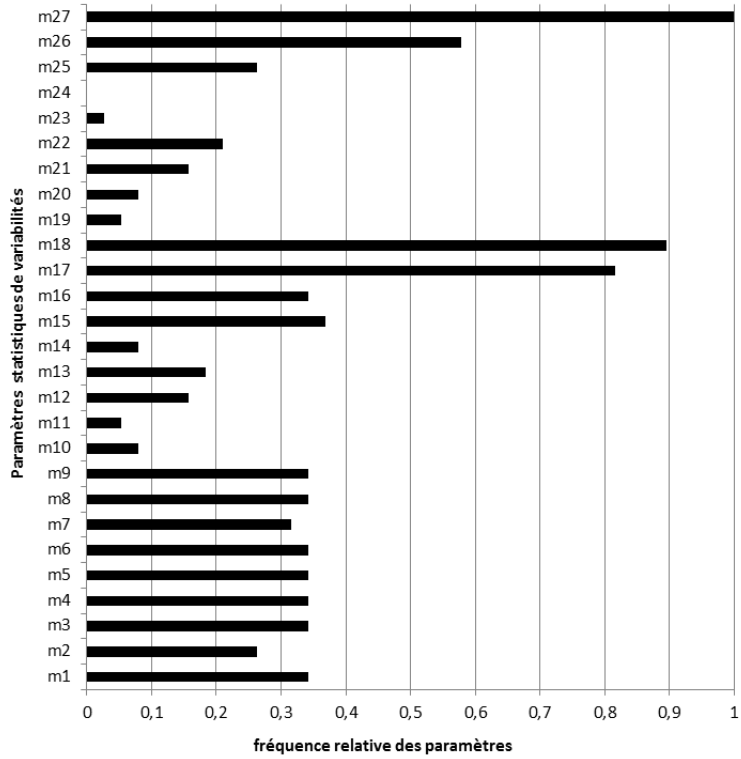


Figure 3: La répartition de la fréquence de sélection des 27 paramètres de variabilités pour une classification optimale en 4 classes.

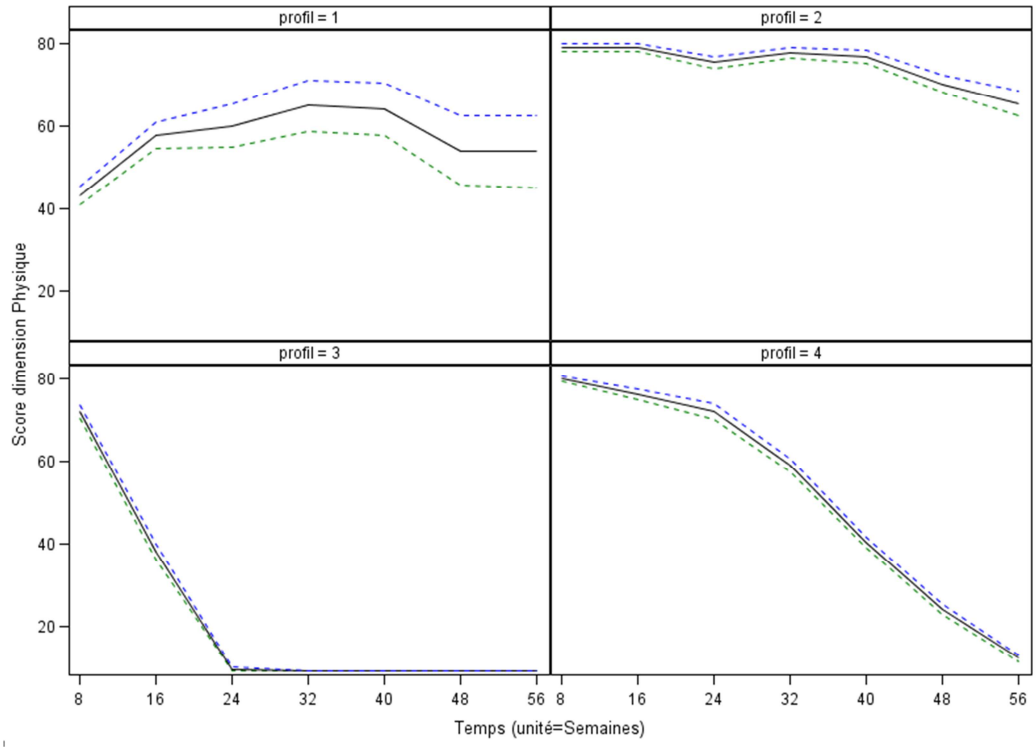


Figure 4: Représentation des profils évolutifs de qualité de vie en fonction du temps

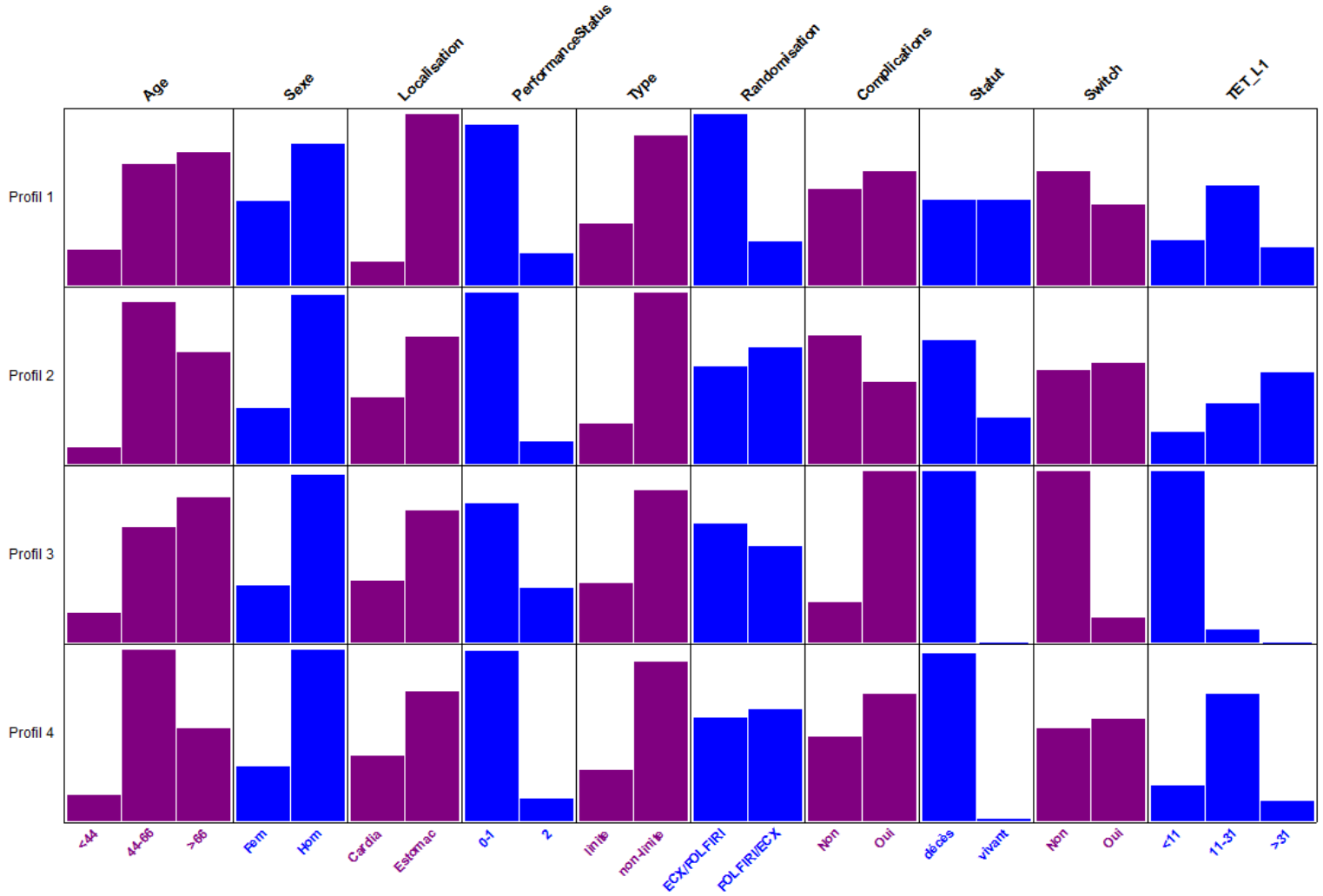


Figure 5: Description des différents profils évolutifs à partir des variables décrivant individuellement chaque patient.

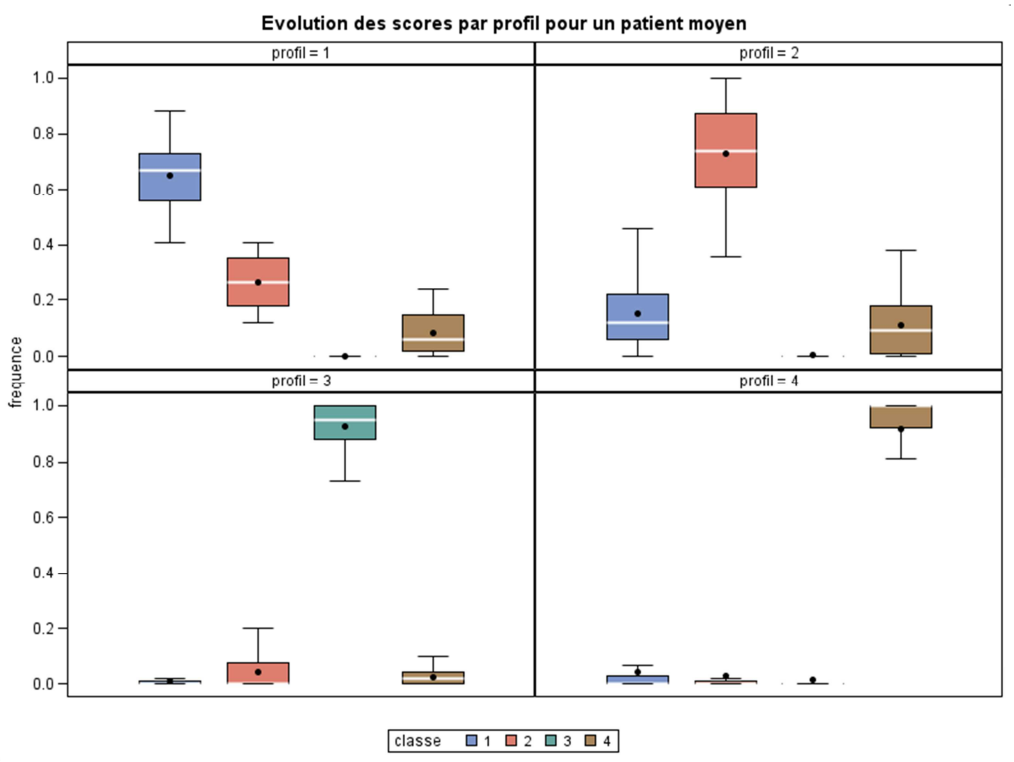


Figure 6: Visualisation pour chaque profil de la représentation des différentes classes d'affectation des profils individuels des scores (PIS) des patients.

VIII. ANNEXES

Annexe 1 : Les étapes de construction des profils évolutifs à partir de données longitudinale de score de qualité de vie et prenant en compte la présence de scores manquants

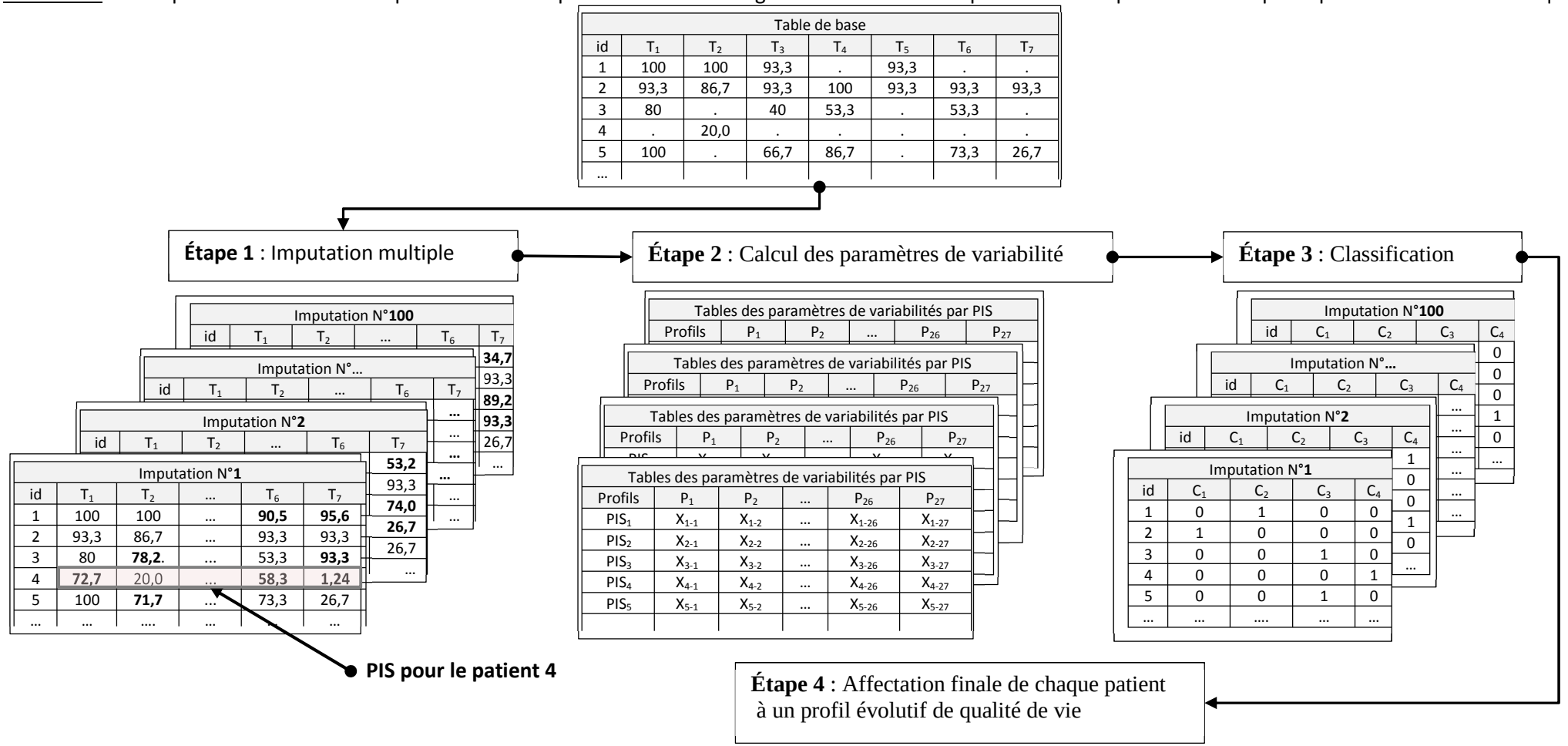


Table de base							
id	T_1	T_2	T_3	T_4	T_5	T_6	T_7
1	100	100	93,3	.	93,3	.	.
2	93,3	86,7	93,3	100	93,3	93,3	93,3
3	80	.	40	53,3	.	53,3	.
4	.	20,0	.	.	.	.	.
5	100	.	66,7	86,7	.	73,3	26,7
...							

Imputation N°1					
id	T_1	T_2	...	T_6	T_7
1	100	100	...	**90,5**	**95,6**
2	93,3	86,7	...	93,3	93,3
3	80	**78,2**.	...	53,3	**93,3**
4	**72,7**	20,0	...	**58,3**	**1,24**
5	100	**71,7**	...	73,3	26,7
...	...	....	...		...

Imputation N°2					
id	T_1	T_2	...	T_6	T_7
					53,2
					93,3
					74,0
					26,7
					26,7
					...

Imputation N°...					
id	T_1	T_2	...	T_6	T_7
					93,3
					89,2
					93,3
					26,7
					...

Imputation N°100					
id	T_1	T_2	...	T_6	T_7
					34,7
					93,3
					89,2
					93,3
					26,7
					...

Tables des paramètres de variabilités par PIS					
Profils	P_1	P_2	...	P_{26}	P_{27}
PIS_1	X_{1-1}	X_{1-2}	...	X_{1-26}	X_{1-27}
PIS_2	X_{2-1}	X_{2-2}	...	X_{2-26}	X_{2-27}
PIS_3	X_{3-1}	X_{3-2}	...	X_{3-26}	X_{3-27}
PIS_4	X_{4-1}	X_{4-2}	...	X_{4-26}	X_{4-27}
PIS_5	X_{5-1}	X_{5-2}	...	X_{5-26}	X_{5-27}

Imputation N°1				
id	C_1	C_2	C_3	C_4
1	0	1	0	0
2	1	0	0	0
3	0	0	1	0
4	0	0	0	1
5	0	0	1	0
...	...	....	...	...

Imputation N°2				
id	C_1	C_2	C_3	C_4
				1
				0
				0
				1
				0
				...

Imputation N°...				
id	C_1	C_2	C_3	C_4
				...
				...
				...
				...
				...

Imputation N°100				
id	C_1	C_2	C_3	C_4
				0
				0
				0
				1
				0
				...

Id : identification unique des patients
PIS : Profil individuel de scores
T_1, T_2, …, T_7 : Temps d'évaluation de la qualité de vie
P_1, P_2, …, P_{26}, P_{27} : Paramètres de variabilité calculés sur chaque PIS
C_1, C_2, …, C_4 : Classe d'affectation à l'issue de la classification
PE : Profil évolutif (ou l'affectation finale) de chaque patient.

id	C_1	C_2	C_3	C_4	PE
1	0,98	0	0,02	0	**1**
2	0	0,24	0,74	0,02	**3**
3	0,12	0,32	0,28	0,28	**2**
4	0,16	0,18	0,44	0,22	**3**
5	0,1	0,32	0,36	0,22	**2**
...	...	....	...	...	...

<u>Annexe 2 :</u> Les différentes formes de distribution des scores manquant dans un profil individuel de scores de patients.

Profil de score complet	Profil de score avec données manquantes		
	Monotone	Intermittente	Mixte
XXXXXXX[1]	XXXX---[2]	XX-XX-X	X-XXX--
XXXXXXX	XXX-----	X--XXXX	XX-X---
XXXXXXX	X------	-XXX--X	----X--
XXXXXXX	XX-----	--XX-XX	-X-X-X-

[1]X= score présent [2]-=score manquant

Annexe 3 : tableau des mesures sélectionnées pour la classification

N°	Mesure sur y_i (i=1,…, k)[1]	Formule
Mesures de linéarité		
1	étendu	max y_i – min y_i
3	Écart-type (SD)	$S_y=\sqrt{\frac{1}{k-1}(y_i-\bar{y})^2}$
4	Coefficient de variation (CV)	$(S_y/\bar{y})\times 100$
5	Changement	y_k-y_1
6	Moyenne du changement par unité de temps	$(y_k-y_1)/(t_k-t_1+1)$
8	Changement relatif à la moyenne dans le temps	$(y_k-y_1)/\bar{y}$
9	Pente du modèle linéaire $y_i=a+bt_i+\varepsilon_i$	$b=\frac{\sum_{i=1}^{k}(y_i-\bar{y})(t_i-\bar{t})}{\sum_{i=1}^{k}(t_i-\bar{t})^2}$
Mesure de non linéarité et de la variation des changements basée sur les différences de 1er niveau ($\Delta_{(1,i)}=y_{(i+1)}-y_i$)		
15	Maximum des différences de 1er niveau absolues	max $\lvert\Delta_{1,i}\rvert$
16	Rapport du maximum des différences de 1er niveau absolues à la moyenne globale	$(\max\lvert\Delta_{1,i}\rvert)/\bar{y}$
17	Rapport du maximum des différences de 1er niveau absolues à la pente	$(\max\lvert\Delta_{1,i}\rvert)/b$
18	Rapport de SD à la pente	(S_{Δ_i}/b)
Mesures de contraste avant/après[2]		
26	Rapport entre la mesure du changement avant et celle globale	$(y_c-y_1)/(y_k-y_1)$
27	Rapport entre la mesure du changement après et celle globale	$(y_k-y_{c+1})/(y_k-y_1)$

[1] k pouvant varier d'un patient à l'autre

[2] Les mesures d'avant/après sont réalisées à partir d'un temps de coupure définie dans les observations

IX. RÉFÉRENCE BIBLIOGRAPHIQUES

1. Bouvier AM, Remontet L, Jougla E, Launoy G, Grosclaude P, Buemi A, et al. Incidence of gastrointestinal cancers in France. Gastroenterol Clin Biol. 2004;28(10 Pt 1):877-81.

2. Remontet L, Esteve J, Bouvier AM, Grosclaude P, Launoy G, Menegoz F, et al. Cancer incidence and mortality in France over the period 1978-2000. Rev Epidemiol Sante Publique. 2003;51(1 Pt 1):3-30.

3. Moinpour C, Donaldson G, Liepa A, Melemed A, O'Shaughnessy J, Albain K. Evaluating health-related quality-of-life therapeutic effectiveness in a clinical trial with extensive nonignorable missing data and heterogeneous response: results from a phase III randomized trial of gemcitabine plus paclitaxel versus paclitaxel monotherapy in patients with metastatic breast cancer. Quality of Life Research. 2012;21(5):765-75.

4. Curran D, Pozzo C, Zaluski J, Dank M, Barone C, Valvere V, et al. Quality of life of palliative chemotherapy naive patients with advanced adenocarcinoma of the stomach or esophagogastric junction treated with irinotecan combined with 5-fluorouracil and folinic acid: results of a randomised phase III trial. Qual Life Res. 2009;18(7):853-61.

5. Sadighi S, Mohagheghi MA, Montazeri A, Sadighi Z. Quality of life in patients with advanced gastric cancer: a randomized trial comparing docetaxel, cisplatin, 5-FU (TCF) with epirubicin, cisplatin, 5-FU (ECF). BMC Cancer. 2006;6:274.

6. Gotay CC. Assessing cancer-related quality of life across a spectrum of applications. J Natl Cancer Inst Monogr. 2004(33):126-33.

7. Young T, deHaes H, Curran D. Guidelines for assessing quality of life in EORTC. 2004.

8. Beitz J, Gnecco C, Justice R. Quality-of-life end points in cancer clinical trials: the US Food and Drug Administration perspective. J Natl Cancer Inst Monogr. 1996;20:7-9.

9. Barbare J, Bouche O, Bonnetain F, Raoul J, Rougier P, Abergel Aea. Randomized controlled trial of tamoxifen in advanced hepatocellular carcinoma. J Clin Oncol. 2005;23(19):4338-46.

10. Bedenne L, Michel P, Bouche O, Milan C, Mariette C, Conroy Tea. Chemoradiation followed by surgery compared with chemoradiation alone in squamous cancer of the esophagus: FFCD 9102. J Clin Oncol. 2007;25(10):1160-8.

11. Giesler RB. Assessing the quality of life in patients with cancer. Curr Probl Cancer. 2000;24(2):58-92.

12. Fairclough DL. Design and analysis of quality of life studies in clinical trials (Second ed.): Chapman and Hall/CRC Press; 2010.

13. Michikazu N, Weimin K. Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. International Journal of Mathematical Analysis. 2011;5(1):1-13.

14. Leffondre K, Abrahamowicz M, Regeasse A, Hawker G, Badley E, McCusker Jea. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. J Clin Epidemiol. 2004;57(10):1049-62.

15. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol. 2010;6:109-38.

16. Shi Q, Mendoza T, Gunn GB, Wang X, Rosenthal D, Cleeland C. Using group-based trajectory modeling to examine heterogeneity of symptom burden in patients with head and neck cancer undergoing aggressive non-surgical therapy. Quality of Life Research. 2013:1-9.

17. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine. 2012;367(14):1355-60.

18. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009;18(1):1-43.

19. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549-76.

20. Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. Am J Epidemiol2013.

21. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. 1995;90(431):1112-21.

22. Vickery CW, Blazeby JM, Conroy T, Arraras J, Sezer O, Koller M, et al. Development of an EORTC disease-specific quality of life module for use in patients with gastric cancer. Eur J Cancer. 2001;37(8):966-71.

23. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365-76.

24. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

25. Edwin D. Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. Revue de Statistique appliquée. 1971;19(2):19-33.

26. Breaban M, Luchian H. A unifying criterion for unsupervised clustering and feature selection. Pattern Recognition. 2011;44(4):854-65.

27. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193-218.

28. Warrens M. On the Equivalence of Cohen’s Kappa and the Hubert-Arabie Adjusted Rand Index. Journal of Classification. 2008;25(2):177-83.

29. Warrens M. A Formal Proof of a Paradox Associated with Cohen’s Kappa. Journal of Classification. 2010;27(3):322-32.

30. Nuemi G, Afonso F, Roussot A, Billard L, Cottenet J, Combier E, et al. Classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy. Cancer Epidemiol. 2013;37(5):688-96.

31. Post WJ, Buijs C, Stolk RP, de Vries EG, le Cessie S. The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. Qual Life Res. 2010;19(1):137-48.

X. PUBLICATIONS

G. Nuemi, H. Devilliers, K.Le Malicot, R. Guimbaud, C. Lepage, C. Quantin : construction of quality of life change patterns: example in Oncology in a phase iii therapeutic trial, Qual Life Res.(*Soumis 2014*)

CONSTRUCTION OF QUALITY OF LIFE CHANGE PATTERNS: EXAMPLE IN ONCOLOGY IN A PHASE III THERAPEUTIC TRIAL

G. Nuemi[1,2], H. Devilliers[1,2], K. Le Malicot[2], R. Guimbaud[3], C. Lepage[1,2], C. Quantin[1,2]

[1] CHRU Dijon, Service de Biostatistique et d'Informatique Médicale (DIM), Dijon, F-21000, France

[2]Inserm, U866, Université de Bourgogne, Dijon, France

[3]Inserm, UMR 1037/CNRS-ERL 5294, Université Toulouse 3,Toulouse, France

Corresponding author: Catherine QUANTIN, Centre Hospitalier Universitaire de Dijon, Service de biostatistique et d'Informatique Médicale (DIM), BP 77908, 21079 Dijon Cedex, France ; catherine.quantin@chu-dijon.fr

Abstract

Introduction

In oncology, the complexity of the analysis of data describing quality of life (QoL) related to health remains a barrier to its routine use by clinicians. The aim of this work was to construct a QoL's change patterns for patients with advanced cancer.

Materials and methods

We used data from a phase III trial that compared two strategies of chemotherapy in advanced gastrointestinal cancer. QoL was assessed by self-report questionnaire: the QLQ-C30. Only the scores of the physical dimension were analyzed. Missing data were managed using multiple imputation technique. Four steps were necessary to identify QoL change patterns and each QoL change patterns was graphically described.

Results

The trial included 416 patients and 1023 questionnaire were collected. 74% of patients were male with a mean age of 62±11 years. About 1650 from the 2912 expected (57%) scores were analyzed. The rate of missing score was 43%. Patients were grouped into four classes and a typology with three change patterns was demonstrated : The improved pattern had the lowest population and a maximum difference (MaxDiff) in scores equal to +25pts between the extremes, the stability pattern included patients with baseline score around 80/100 and MaxDiff=-12pts. This was the most provided patient's pattern (41%). Finally, the degradation pattern organised on 2 groups with the main difference due to the initial scores: 43/100 versus 82/100. These patterns were graphically represented using frequency bar charts.

Conclusion

This work opens up perspectives for longitudinal data analysis with high probability of missing values while providing a relevant graphical summary.

Keywords

Quality of life; change patterns; multiple imputation; Clustering

INTRODUCTION

Gastrointestinal cancers are among the most frequent cancers in France (1, 2). Metastatic or locally-advanced cancers have a bleak prognosis. The treatments proposed only marginally improve survival, but allow progress in terms of quality of life (QoL) (3-5). QoL is recognized as an essential criterion for the evaluation of new treatments (6-8) and a description of the evolution of patients' feelings in the context of aggressive therapy is becoming a systematic secondary objective in phase III trials, and in a large number prognostic studies in routine practice (9, 10).

However, the analysis of QoL in patients is a very complex process and the heterogeneity of patients' perception of their state of health often makes it difficult for clinicians to interpret the objective results of studies. We can first of all underline the subjective nature of the information, which, in addition, evolves with time. Then, collecting the data currently requires the use of multidimensional measurement scales. Finally, the longitudinal nature of the data makes it almost impossible to avoid missing data (MD); the reasons may be wide and varied (11). The impact of MD, given the variable proportions, must be taken into account to attenuate the non-negligible risk of bias (11).

Nonetheless, taking this subjective element into account is particularly important with regard to the results of phase III therapeutic trials involving patients with advanced gastrointestinal cancer for whom the objective to improve quality of life is essential. Despite its complexity, the evolution of QoL must be described, presented and interpreted in a way that is simple and understandable. The methods used to analyze longitudinal data have been evolving steadily over the last twenty years or so and notably since the use descriptive methods, which use measurements of central trends (means, medians,…) and thus eclipse the longitudinal nature of the data. We can note, for example, the contribution of hierarchical models, which allowed a more precise description of the data but for which the results remained difficult to interpret for the non-specialist (12). Today, even more sophisticated methods make it possible to identify subgroups (clusters) of QoL change patterns from quantitative variables and elaborate specific hypothesis in each subgroup of patients (13-15). These methods are becoming easier to use and provide easily interpretable results. At the same time, methods to take MD into account in the analysis of clinical studies are being standardized (12, 16-20).

The aim of this work was to use these new methods firstly to construct a typology of QoL change patterns (CP) in the context of a phase III therapeutic trial in patients with locally-

advanced and metastatic gastrointestinal cancer, and secondly to describe identified patterns using the variables collected.

MATERIALS AND METHODS

Study design

We worked on data from a phase III clinical trial (FFCD-03-07) conducted between June 2005 and May 2010, the design of which is described elsewhere. It was a multicenter, randomized, open, prospective trial that compared the efficacy of two courses of polychemotherapy (FOLFIRI versus Epirubicin-Cisplatin-Capecitabine (ECX) and vice-versa) in patients with locally-advanced or metastatic adenocarcinoma of the stomach or the cardia. This trial showed that time to treatment failure with ECX as the first-line therapy (TTF) was significantly shorter than that with first-line FOLFIRI.

Collection of quality of life data

The secondary objective was to evaluate quality of life measured by the EORTC-QLQ-C30 (European Organisation for Research and Treatment of Cancer- Quality of life questionnaire core 3.0) self-administered questionnaire. The study design provided for an evaluation every 8 weeks using the QoL self-administered questionnaire.

EORTC QLQ-C30 Questionnaire

The QLQ-C30 is a scale of measurements with 30 items that covers health issues that are relevant to a wide range of cancer patients. Of those 30 items, 17 are grouped into 5 functional scales or dimensions (physical, cognitive, role, emotional, and social functioning) and one global health status/quality of life scale. The remaining 13 items are grouped into scales measuring cancer related symptoms (fatigue, nausea/vomiting, pain, dyspnea, diarrhea, insomnia, appetite loss, and constipation). The questionnaire is validated in patients with gastric cancers (21). For each scale, a score is calculated according to a standardized method (22). The functional and global health status scale scores can range from 0 (severe debilitation) to 100 (asymptomatic/best quality of life) and the symptom scale scores can range from 0 (asymptomatic) to 100 (severe debilitation). We studied the physical functional scale based on 5 items (question 1 to the 5^{th}). For patients, it means assessing for the previous week their abilities to carry out certain everyday activities such as getting dressed, taking care of personal hygiene, carrying a bag of shopping or even going for a walk.

From individual score patterns to QoL change patterns

The first seven measurements of the physical functional scale score (QLQ-C30 PF2) included in the protocol (every 8 weeks following the date of randomization), corresponding to a follow-up of approximately 14 months, were analyzed. Hereinafter, they will be noted T_1, T_2, …, T_7. An individual score patterns (ISP) was defined as the chronological series of the different scores calculated at each time point for a given patient. Starting with the initial table of ISP, four steps were necessary to identify quality of life change patterns using these variables. These steps are resumed in annex 2:

1/The first step was a process of data augmentation.(17, 23) In the context of our study, the presence of MD could make it impossible to calculate some of the 27 parameters. To circumvent this caveat, we applied the multiple imputation (MI) method on scores. This method may be used for the analysis of data with large amounts of missing values (12, 16-20). One hundred datasets with complete ISP were generated. This method finally encompassed the major part of patient's quality of life measures variability. For the main analysis, scores after death were set to zero.

2/In the second step, some variability parameters were calculated from each of the 100 imputed datasets. These parameters were proposed by Leffondre et al. in 2004 for the identification of longitudinal patterns (13). There were 27 simple statistical parameters: i) parameters that described the linearity of the ISP (e.g. the standard deviation, the slope of the regression line or the part of the variance explained by a linear model), ii) those that reflected non-linearity of the ISP such as abrupt changes over short periods (e.g. the mean of successive differences between 2 consecutive scores), iii) parameters that measured the contrast between 2 defined periods in an ISP (ratio between the change before and the change after). A sample of these parameters is shown in annex 1. Each ISP was thus described by 27 parameters in each of the 100 datasets from the imputation. At this stage, we had 100 new datasets that included the variability parameters for each ISP.

3/The third step aimed at building subgroups (clusters) of ISP, applying a classification method to each dataset that included the variability parameters. A non-supervised classification technique based on the « k-means» method, using Euclidian distances was applied (19, 24). The number of clusters and the variability parameters used for the classification were automatically selected by maximizing a so called CritCF criteria (ranged from 0 to 1) (25) with a standardized methodology (19, 24, 25).

4/The last step was to classify each patient in the group that best reflected his own QoL change pattern. This objective was reached by the mean of the aggregation of the 100 classifications' results. The aggregation process was carried out in accordance with a published methodology (19). A given patient was assigned to the change pattern (CP) in which he was most frequently classified within the 100 classifications. For each CP, a mean ISP was calculated and represented.

Analysis of sensitivity

The objective here was to evaluate the impact of the multiple imputation procedure on the final classification of a patient in one of the CP. For this we initially for each CP, presented graphically as a box plot, the probability of assignment of patients in one of the classes from the clustering process on each of the 100 datasets. This could allow one to observe for example if the patient assigned to the first CP were mainly grouped in the first cluster. In a second time we estimated a coefficient of concordance (Adjusted Rand index) between assignments in different clusters and final assignment to the corresponding CP. This coefficient was presented in the form of a confidence interval of 95% of the mean value (26-28).

Graphical description of QoL change patterns

Each QoL change patterns was described using patients' parameters: Individual data (age and sex), randomization data (tumors type and location, randomization arm and World Health Organization (WHO) Performance Status), and follow-up data (TTF, declaration of serious adverse events (Yes/no), second line treatment administration (Switch Yes/no) and the status of the patient (Alive or dead). These parameters were presented in the form of frequency bar charts where each bar corresponded to a parameter modality and the height was its relative frequency(29) .

Frequencies were compared using the Chi-2 test or the Fisher's exact test when appropriate. Survival rates were estimated using the Kaplan-Meier method. The CP were compared using the difference the maximum and the minimum score (MaxDiff). For all of the statistical tests, the threshold of statistical significance was set at 0.05.

SAS version 9.3. was used for all of the statistical analyses

RESULTS

The clinical trial had included 416 evenly randomized patients (209 in ECX and 207 in FOLFIRI arm, respectively). Men accounted for the majority of patients in both arms (approximately 74%). The death rate in both arms after 14 months of follow-up was approximately 55 %. The TTF observed were greater in the FOLFIRI/ECX arm (30). Table 1 presents the characteristics of patients.

The 1023 exploitable self-administered questionnaires came from 364 patients, that is a mean of 3±2 (range [1; 12]) questionnaires per patient. Of the first 7 evaluations, 2 912 self-administered questionnaires (one score for the physical dimension) were expected given the total number of patients who had received at least one dose of one treatment (416). Considering that for the patients who had died, the absence of a score after the death was not MD, we counted 1 262 missing scores, that is a proportion of MD of approximately 43%.

With the process of maximizing the CritCF criteria, we retain the classification in 4 clusters and 13 variability parameters listed in annex 1. These statistical parameters were selected as the most frequent as shown in figure 1. The corresponding CritCF mean value was 0.75 ranging from 0.61 to 0.87.

The 416 patients were definitively assigned to one of the 4 QoL change patterns P1, P2, P3 and P4 in the following proportions: 6%, 41%, 19% and 34%, respectively. Considering the MaxDiff value, we described a typology in 3 patterns: an improving pattern (P1) with MaxDiff=+25 pts, a stability pattern P2 (MaxDiff=-12pts) and 2 deterioration patterns P3 (MaxDiff=-21pts) and P4 (MaxDiff=-27pts). Figure 2 showed the trend for the mean scores per QoL change patterns over the 7 evaluation time points. These trend curves were bordered with curves showing the standard deviation of the mean scores. Each QoL change patterns was summarized using classical statistics computed from patient available data (age, sex, randomization arm, performance status, observed death rate and incidence of serious adverse events) as shown in table 2. Note that the linear trend of each pattern was recalled in this table header.

Figure 3 showed a much more visual description of the different patterns according to certain variables. The TTF variable made it possible to distinguish between the 4 QoL change patterns. By taking into account information from table 2, the 4 patterns could be distinguished one from the other as follows: Pattern 1 mostly concerned with patients from the ECX/FOLFIRI arm, in pattern 2, we had relatively younger patients (age between 44 and

66 years old) with the highest TTF (above 31 weeks); for pattern 3, the majority of patients experienced serious adverse events and had the lowest TTF (less than 11 weeks) and pattern 4 essentially comprised patients with a TTF between 11 and 31 weeks, but had the highest number of deaths.

The impact of the distribution of imputed values is represented in figure 4, which shows the distribution of the initial ISP clusters for each change pattern. Pattern 2, for example, comprised a majority of patients from cluster 2 (73% on average), but also patients from cluster 1 at 15% and very rarely patients from cluster 3. In the same way, the concordance coefficient between each of the 100 classification and the final assignment was 0.62 ($_{95\%}$ CI [0.61-0.63]) on average.

DISCUSSION

This work allowed us to show that it was possible to construct pertinent QoL change patterns for patients using longitudinal QoL scale measurements. As a result, 3 types of pattern were described: the improving pattern (P1), which had the smallest number of patients and a maximum difference (MaxDiff) between scores of +25 points; the stability pattern (P2), which included patients with a mean initial score equal to 80 and MaxDiff=-12 pts – this was the pattern with the most patients (41%); and finally, two deterioration patterns distinguished according to the patients' initial score: in the first group (P3), with the smallest number of patients (19%), the initial QLQ-C30 physical score was around 43 while in the second (P4), it was around 82 on the 0-100 scale. And finally, we have proposed a graphical representation of these patterns. For the variables collected at inclusion, such as the location and type of tumor, the Performance status and sex, we found that the structure of the bar chart was similar whatever the pattern. However, for the variable, age, the structure was nuanced with a predominance of the oldest patients in patterns P1 and P3, and, relatively young patients were predominant in patterns P2 and P3. For variables collected throughout the study, such as serious adverse events (complications) or the time to therapeutic failure of the 1st line therapy (TTF), the structure of the bar chart varied from one pattern to another.

Most studies on QoL in cancer focus on the effects of a given treatment on QoL. Our approach was different in that it aimed to define the clinical profile associated with a given pattern of QoL evolution. These clinical profiles may have a clinical impact because they could help clinicians to predict the, the course of QoL during cancer treatment. For example, clinicians could use such profiles to identify patients with a low performance status at baseline but who could present a deterioration in their QoL by looking, for example, for

adverse events. Patterns P3 and P4, for example, which include patients with a low performance status at baseline and the occurrence of serious adverse effect (SAE), were associated with a physical QoL pattern of deterioration and a low baseline score. On the other hand, clinicians may wish to identify patients who could present an improvement in their QoL. For example, clinicians could select patients in the ECX arm with TTF mainly between 11 and 31 weeks as, in our study, we showed that patients classified in pattern P1 presented an improvement in their QoL compared with the other patterns. Clinicians may also want to know which patients will not present a major change in their QoL. In this case, they would be interested in selecting patients with TTF above 31 weeks (whatever the treatment arm) classified in pattern P2, which showed relative stability in our study.

Our results are in coherence with the literature. For example, Sadighi et al. (5) also found that the patients who expressed a deterioration in their quality of life (pattern P3 in our study) also presented the shortest TTF (<11 weeks) and had been given ECX as the first-line treatment (ECX/FOLFIRI arm). In another paper, Curran et al (4) also showed that patients who do not present a major change in their QoL (pattern P2 of stability in our study) have good clinical results (TTF above 31 weeks) and were principally given irinotecan as the first-line treatment (FOLFIRI/ECX arm).

It seems important to point out that these results should not eclipse the strategy used for missing data. The principal advantage of using the multiple imputation method is the conservation of data distribution. In our study, this strategy led to a model that allows a more true-to-life interpretation of reality, as it is less biased than the model we would have obtained using other imputation methods (using means, last available value, etc). Nonetheless, this model still contains a certain number of pitfalls inherent to the construction of any statistical model. The advantage of this method we used to construct the patterns was the fact that it produced patterns that were a priori independent of other clinical or personal variables collected. This means that it would be possible to analyze associations between these patterns and the study variables (31) or to search for predictors of these patterns. Certain steps in the implementation of this method, such as the choice of the number of patterns could be automated, so that this process would not turn out to be rather fastidious. As a graphical representation makes it easier for clinicians to interpret the data, the presentation we proposed here (bar charts) makes it possible to visualize (29) the variability between the different

patterns sometimes difficult to grasp when it is reduced to standard deviations or confidence intervals alone.

CONCLUSION

The aim of this work was to show that clinical trial results may be used to identify clinical profile associated with QoL change over the time. To achieve this, we purposely chose an unfavorable situation, a complex therapeutic trial in patients with severe, advanced cancer and thus a high foreseeable risk of missing values for quality of life data. Thanks to the rigorous methodology, the results suggest the interest of conducting deeper analyses (notably by studying the sensitivity of the results), of extending the analyses to other scales of the EORTC-QLQ-C30 questionnaire and of comparing the results with data from other therapeutic trials. Perspective for future research includes conducting the analysis on other EORTC QLQ-C30 scales, pooling trials to obtain general QoL evolution patterns, and comparing QoL evolution pattern according to detailed socio-demographic and clinical profile to help clinician identifying *a priori* patients that are likely to have a degradation of QoL pattern.

ACKNOWLEDGEMENTS:

This research was funded by the regional council of Burgundy.

CONFLICT OF INTEREST: none

REFERENCES:

1. Bouvier AM, Remontet L, Jougla E, Launoy G, Grosclaude P, Buemi A, et al. Incidence of gastrointestinal cancers in France. Gastroenterol Clin Biol. 2004;28(10 Pt 1):877-81.
2. Remontet L, Esteve J, Bouvier AM, Grosclaude P, Launoy G, Menegoz F, et al. Cancer incidence and mortality in France over the period 1978-2000. Rev Epidemiol Sante Publique. 2003;51(1 Pt 1):3-30.
3. Moinpour C, Donaldson G, Liepa A, Melemed A, O'Shaughnessy J, Albain K. Evaluating health-related quality-of-life therapeutic effectiveness in a clinical trial with extensive nonignorable missing data and heterogeneous response: results from a phase III randomized trial of gemcitabine plus paclitaxel versus paclitaxel monotherapy in patients with metastatic breast cancer. Quality of Life Research. 2012;21(5):765-75.
4. Curran D, Pozzo C, Zaluski J, Dank M, Barone C, Valvere V, et al. Quality of life of palliative chemotherapy naive patients with advanced adenocarcinoma of the stomach or esophagogastric junction treated with irinotecan combined with 5-fluorouracil and folinic acid: results of a randomised phase III trial. Qual Life Res. 2009;18(7):853-61.
5. Sadighi S, Mohagheghi MA, Montazeri A, Sadighi Z. Quality of life in patients with advanced gastric cancer: a randomized trial comparing docetaxel, cisplatin, 5-FU (TCF) with epirubicin, cisplatin, 5-FU (ECF). BMC Cancer. 2006;6:274.
6. Gotay CC. Assessing cancer-related quality of life across a spectrum of applications. J Natl Cancer Inst Monogr. 2004(33):126-33.
7. Young T, deHaes H, Curran D. Guidelines for assessing quality of life in EORTC. 2004.
8. Beitz J, Gnecco C, Justice R. Quality-of-life end points in cancer clinical trials: the US Food and Drug Administration perspective. J Natl Cancer Inst Monogr. 1996;20:7-9.
9. Barbare J, Bouche O, Bonnetain F, Raoul J, Rougier P, Abergel Aea. Randomized controlled trial of tamoxifen in advanced hepatocellular carcinoma. J Clin Oncol. 2005;23(19):4338-46.
10. Bedenne L, Michel P, Bouche O, Milan C, Mariette C, Conroy Tea. Chemoradiation followed by surgery compared with chemoradiation alone in squamous cancer of the esophagus: FFCD 9102. J Clin Oncol. 2007;25(10):1160-8.
11. Fairclough DL. Design and analysis of quality of life studies in clinical trials (Second ed.): Chapman and Hall/CRC Press; 2010.
12. Michikazu N, Weimin K. Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. International Journal of Mathematical Analysis. 2011;5(1):1-13.
13. Leffondre K, Abrahamowicz M, Regeasse A, Hawker G, Badley E, McCusker Jea. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. J Clin Epidemiol. 2004;57(10):1049-62.
14. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol. 2010;6:109-38.
15. Shi Q, Mendoza T, Gunn GB, Wang X, Rosenthal D, Cleeland C. Using group-based trajectory modeling to examine heterogeneity of symptom burden in patients with head and neck cancer undergoing aggressive non-surgical therapy. Quality of Life Research. 2013:1-9.
16. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The Prevention and Treatment of Missing Data in Clinical Trials. New England Journal of Medicine. 2012;367(14):1355-60.
17. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009;18(1):1-43.

18. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549-76.
19. Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. Am J Epidemiol2013.
20. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. 1995;90(431):1112-21.
21. Vickery CW, Blazeby JM, Conroy T, Arraras J, Sezer O, Koller M, et al. Development of an EORTC disease-specific quality of life module for use in patients with gastric cancer. Eur J Cancer. 2001;37(8):966-71.
22. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365-76.
23. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92.
24. Edwin D. Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. Revue de Statistique appliquée. 1971;19(2):19-33.
25. Breaban M, Luchian H. A unifying criterion for unsupervised clustering and feature selection. Pattern Recognition. 2011;44(4):854-65.
26. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193-218.
27. Warrens M. On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. Journal of Classification. 2008;25(2):177-83.
28. Warrens M. A Formal Proof of a Paradox Associated with Cohen's Kappa. Journal of Classification. 2010;27(3):322-32.
29. Nuemi G, Afonso F, Roussot A, Billard L, Cottenet J, Combier E, et al. Classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy. Cancer Epidemiol. 2013;37(5):688-96.
30. Guimbault R, Louvet C, Ries P, al e. Prospective, randomized, multicenter, phase III study of FOLFIRI vs. ECX in advanced gastric adenocarcinoma: a French intergroup (FFCD-Unicancer-GERCOR) study. J Clin Oncol (In Press). 2014.
31. Post WJ, Buijs C, Stolk RP, de Vries EG, le Cessie S. The analysis of longitudinal quality of life measures with informative drop-out: a pattern mixture approach. Qual Life Res. 2010;19(1):137-48.

TABLES

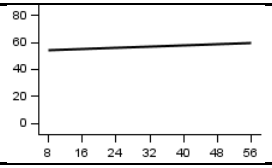
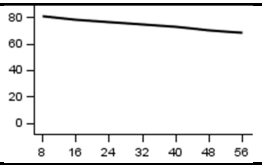
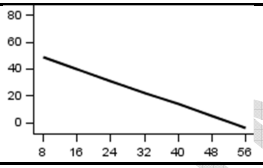
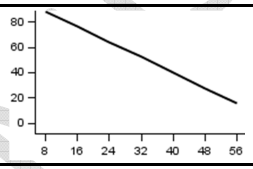
Table 13: Characteristics of patients included in the study

Variables	ECX[1]/FOLFIRI arm ***n=209***	FOLFIRI/ECX arm ***n=207***
Sex		
Men n(%)	154 (74)	155 (75)
Performance status at D_0		
=0-1 n(%)	175(83.7)	178(86.0)
=2 n(%)	34(16.3)	29(14.0)
Type of tumor		
Diffuse n(%)	46(22.0)	51(24.6)
Age (years)		
m±sd[2] [min-max]	61±11 [28-84]	61±11 [29-81]
Follow-up (months)		
m±sd[2] [min-max]	11±10 [0-52]	11±8 [0-42]
SAE[3] n(%)		
After 7 evaluation time points	122 (58)	105 (51)
TTF[4] (weeks)		
m±sd (max)	18±14 (64)	25±23 (137)
Deaths n(%)		
Global[5]	175 (84)	180 (87)
After 7 evaluation time points	116 (56)	113 (55)

[1] *ECX: Epirubicin-Cisplatin-Capecitabine*
[2] *m±sd: mean ± standard deviation*
[3] *Serious adverse events*
[4] *time to therapeutic failure of the first-line treatment*

[5] *Proportion of deaths whatever the therapeutic line*

Table 14: Characteristics of patients' QoL change patterns

	Pattern 1	Pattern 2	Pattern 3	Pattern 4	
					P[1]
Number of patients	24	171	78	143	
Sex					
Men n(%)	15 (63)	128 (75)	58 (74)	108 (76)	0.596
Age (years)					
md±inq[2]	63±17	61±15	65±19	60±17	0.370
Performance status at D_0					
=0-1 n(%)	20(83.3)	151(88.3)	56(71.8)	126(88.1)	0.004
=2 n(%)	4(16.7)	20(11.7)	22(28.2)	17(11.9)	
Type of tumor					
Diffuse n(%)	7(29.2)	33(19.3)	22(28.2)	35(24.5)	0.369
Randomization arm					
FOLFIRI n(%)	5(20.8)	93(54.4)	35(44.9)	74(51.8)	0.015
Quality of life Score					
$Score_7$-$Score_1$ (p^*)	21.8(0.520)	13.5(0.012)	-62.7(0.038)	-67.6($<10^{-3}$)	
SAE[3]					
n(%)	13 (54)	67 (39)	62 (81)	86 (60)	$<10^{-3}$
TTF[4] (weeks)					
md±inq	15±18	30±24	4±7	20±12	$<10^{-3}$
Death					
n(%)	3 (13)	18 (11)	78 (100)	131 (92)	$<10^{-3}$

[1] *: p for significance*
[2] *:median± inter-quartile range*
[3] *Serious adverse events*
[4] *time to therapeutic failure of the first-line treatment*
[*] *: p for significance of the slope predicted by a linear model*

FIGURES

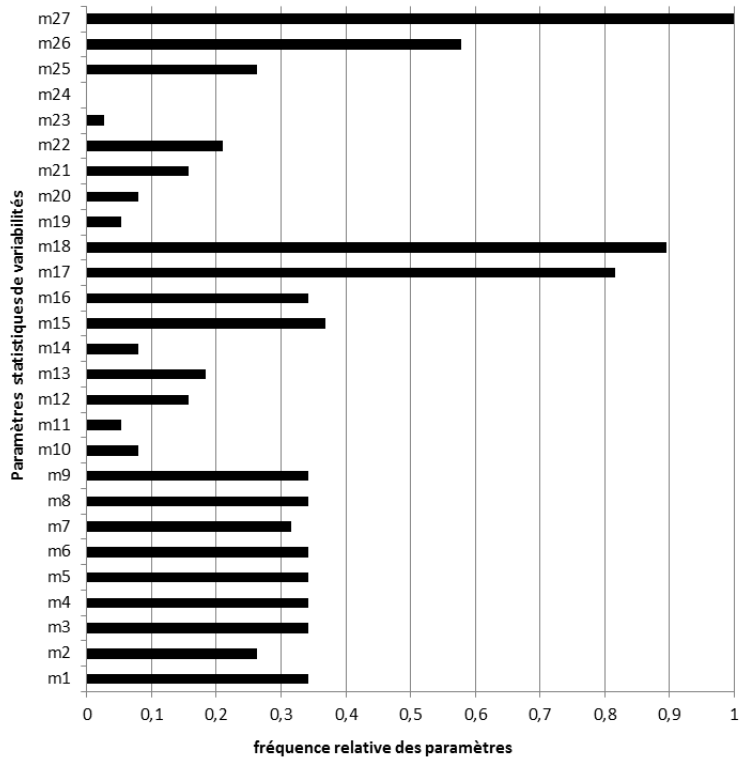


Figure 7: Distribution of the frequencies for the 27 parameters of variability for optimal classification in 4 patterns.

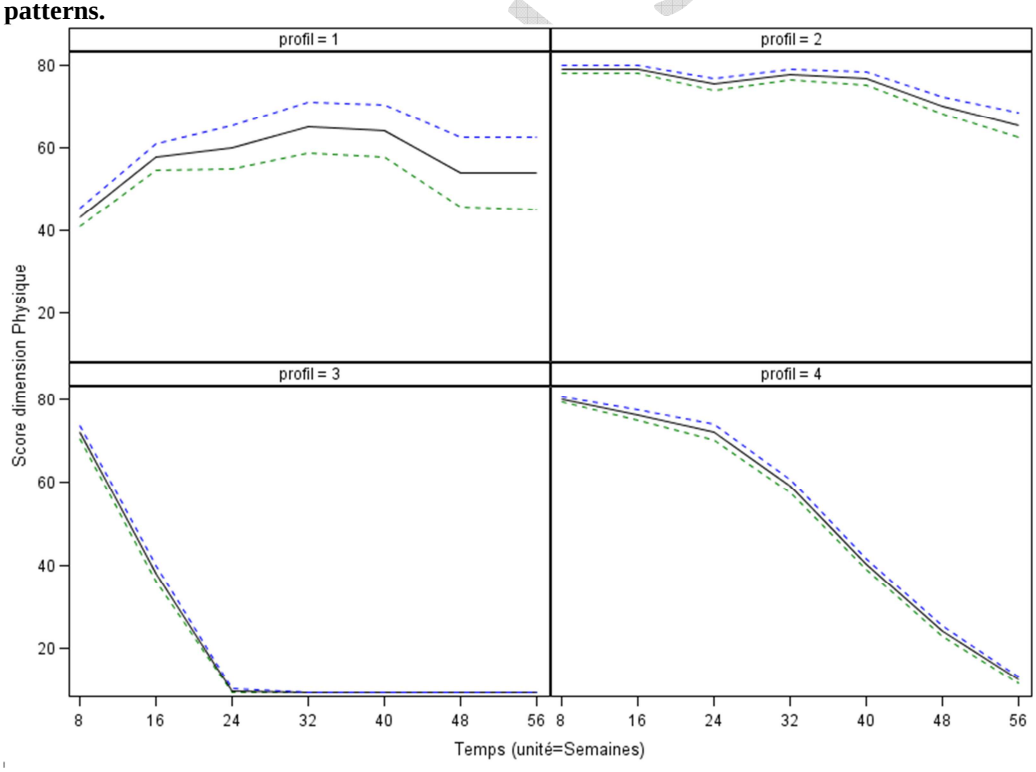


Figure 8: Representation of the quality of life change patterns over time

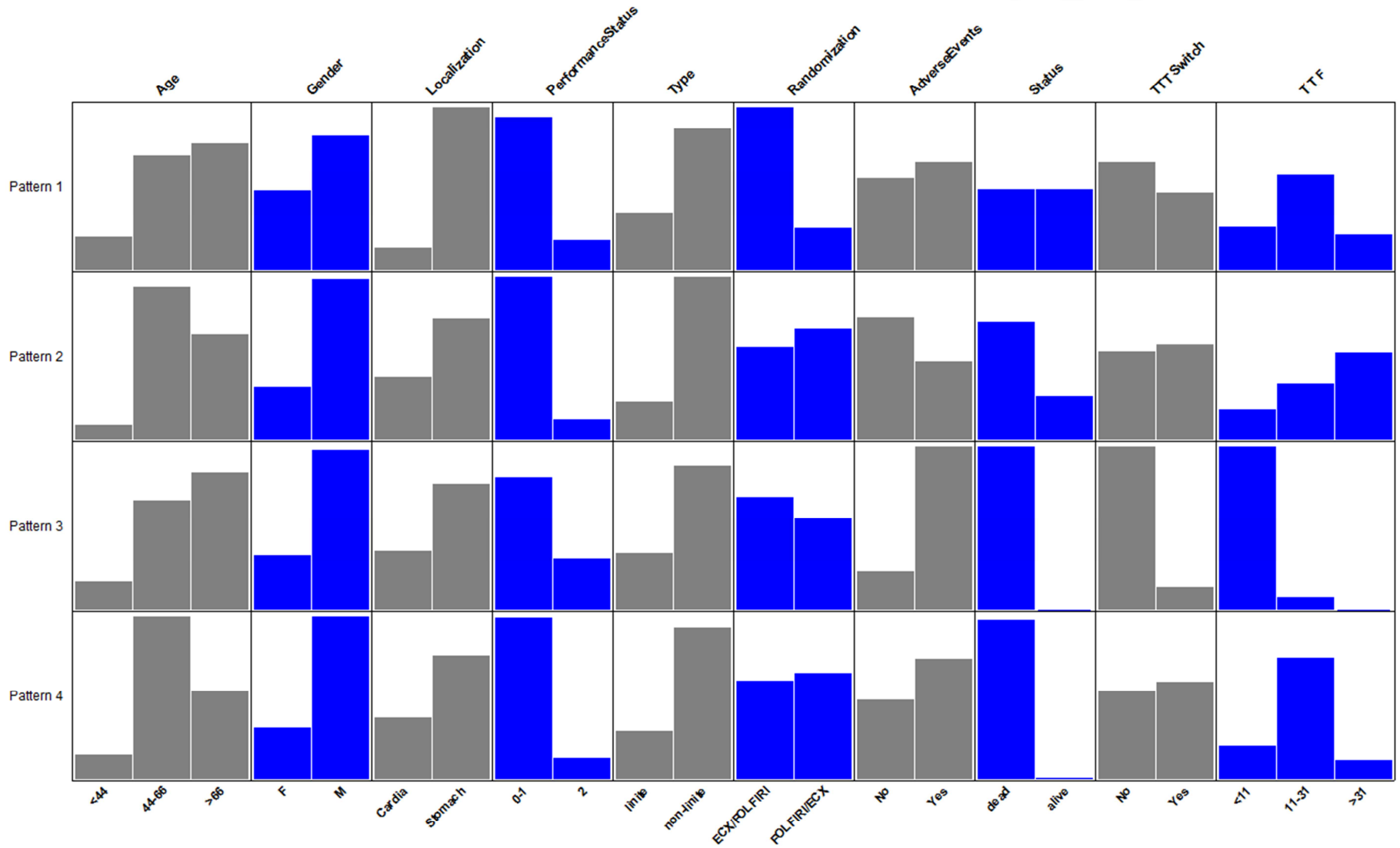


Figure 9: Description of the different change patterns using variables describing each patient individually. ***(Note:staining histogram bar is irrelevant)*** **.**

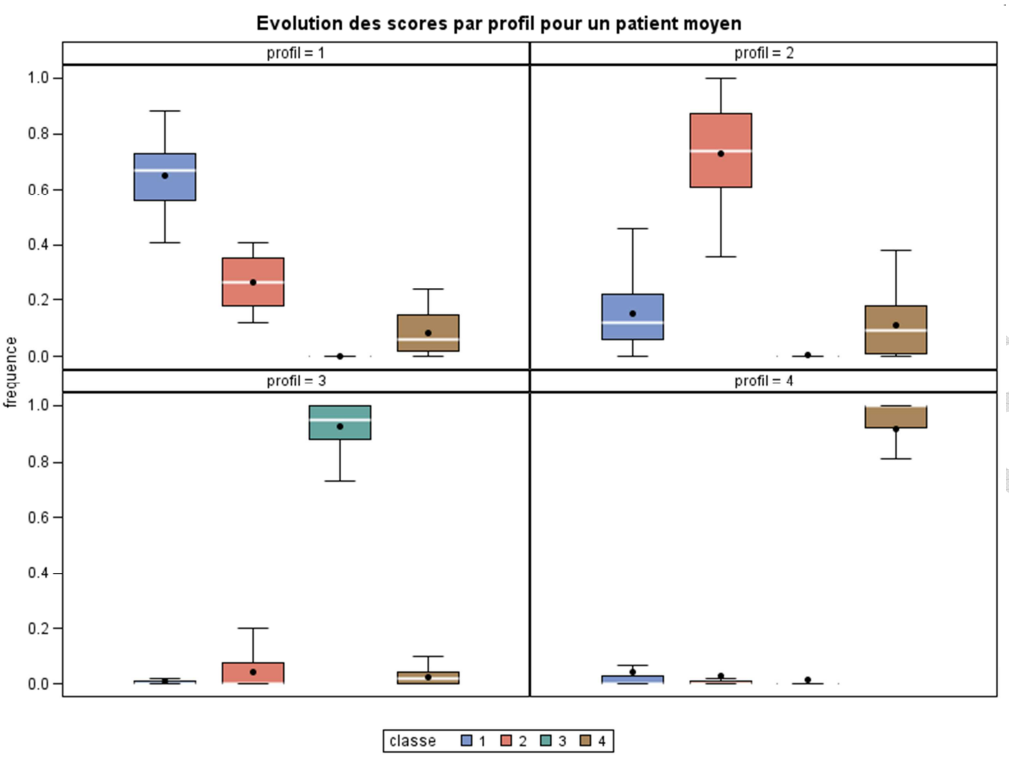


Figure 10: Visualisation for each pattern with regard to the different clusters of patients' individual score patterns (ISP).

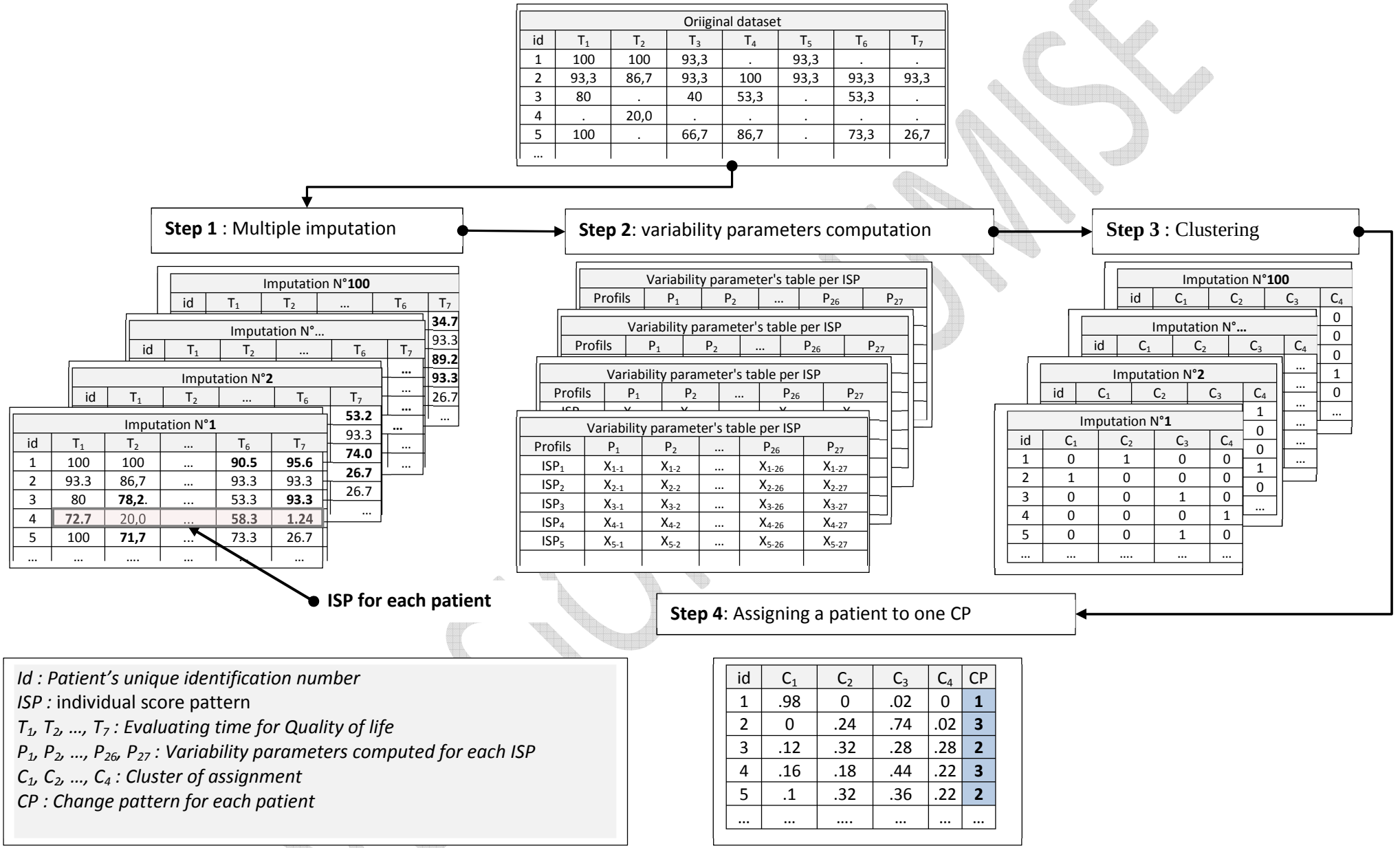
Annex 1: table of measurements selected for the classification

N°	Measurement on y_i (i=1,…, k)[1]	Formula
Measurements of linearity		
1	spread	max y_i – min y_i
3	Standard deviation (SD)	$S_y=\sqrt{\frac{1}{k-1}(y_i-\bar{y})^2}$
4	Coefficient of variation (CV)	$(S_y/\bar{y})\times 100$
5	Change	y_k-y_1
6	Moyenne du change par unite de temps	$(y_k-y_1)/(t_k-t_1+1)$
8	Change relative to the mean over time	$(y_k-y_1)/\bar{y}$
9	Slope of the linear model $y_i=a+bt_i+\varepsilon_i$	$b=\frac{\sum_{i=1}^{k}(y_i-\bar{y})(t_i-\bar{t})}{\sum_{i=1}^{k}(t_i-\bar{t})^2}$
Measurement of non- linearity and variation in changes based on 1st level differences ($\Delta_{(1,i)}=y_{(i+1)}-y_i$)		
15	Maximum absolute differences of the 1st level	max $\lvert\Delta_{1,i}\rvert$
16	Relationship between the maximum absolute differences of the 1st level and the global mean	$(\max\lvert\Delta_{1,i}\rvert)/\bar{y}$
17	Relationship between the maximum absolute differences of the 1st level and the slope	$(\max\lvert\Delta_{1,i}\rvert)/b$
18	Relationship between the SD and the slope	(S_{Δ_i}/b)
Measurements of the before/after contrast[2]		
26	Relationship between the change measured before and global change	$(y_c-y_1)/(y_k-y_1)$
27	Relationship between the change measured after and global change	$(y_k-y_{c+1})/(y_k-y_1)$

[1] k could vary from one patient to another

[2] the before/after measurements were done using a defined cut-off time in the observations

Annex 2:



id	C1	C2	C3	C4	CP
1	.98	0	.02	0	**1**
2	0	.24	.74	.02	**3**
3	.12	.32	.28	.28	**2**
4	.16	.18	.44	.22	**3**
5	.1	.32	.36	.22	**2**
…	…	….	…	…	…

Figure 11: Construction steps of quality of life change pattern from longitudinal data score, taking into account the presence of missing scores

TRAVAUX ANNEXES

Disponible en ligne sur SciVerse ScienceDirect www.sciencedirect.com
Elsevier Masson France EM|consulte www.em-consulte.com
Revue d'Épidémiologie et de Santé Publique
Epidemiology and Public Health
Revue d'Épidémiologie et de Santé Publique 61 (2013) 455–461

Article original

État des lieux et évaluation de la surveillance des *Staphylococcus aureus* résistants à la méticilline (SARM) : PMSI versus surveillance Raisin

Comparing results of methicillin-resistant Staphylococcus aureus (MRSA) surveillance using the French DRG-based information system (PMSI)

G. Nuemi [a], K. Astruc [b], S. Aho [b], C. Quantin [a,*,c]

[a] *Service de biostatistique et d'informatique médicale, CHU de Dijon, CHRU, 21000 Dijon, France*
[b] *Service d'épidémiologie et d'hygiène hospitalière, CHU de Dijon, BP 1519, 21033 Dijon, France*
[c] *Inserm, U866, université de Bourgogne, 21000 Dijon, France*

Reçu le 6 avril 2012 ; accepté le 4 avril 2013

Abstract

Background. – The surveillance of methicillin-resistant *Staphylococcus aureus* (MRSA) is a national priority. The rate of MRSA infections is one of six indicators tracked by the Department of Health. Since 2002, the French institute for public health surveillance (InVS) has monitored MRSA infections to estimate incidence density. Today, the use of the French administrative database (PMSI) could facilitate this surveillance. The aim of this study was to compare MRSA incidence density computed at a national level using PMSI databases with the results from the InVS taken as the reference.

Methods. – PMSI databases for the years 2006 to 2009 were used. The reference results were those published by the InVS from 2006 to 2009. MRSA density defined as the number of MRSA infections recorded per year over 1000 hospitals stays was computed. It was then compared with the MRSA incidence density measured by InVS. The time course of MRSA incidence in the PMSI records was modeled using a Poisson regression.

Results. – The incidence density measured by the InVS was higher than the MRSA density computed using the PMSI, but this difference appeared to decrease over time. The PMSI density/InVS MRSA incidence density ratio was 0.8% in 2006 and about 9.2% in 2009. We observed inverted trends with a growing trend in MRSA density identified by the PMSI. Furthermore, the year of study was significantly associated with incidence density ($P = 0.01$).

Conclusion. – Using PMSI data as an additional source of information in the hospital MRSA surveillance process makes it possible to detect and analyze patient repeats at the regional and national levels with linkage facilities. Estimation of incidence density for hospitals not participating to this surveillance system will be the next step.

Keywords: Methicillin-resistant *Staphylococcus aureus* (MRSA); Cross infection; Diagnosis related groups (DRG); Hospital information systems; Regression analysis; Population surveillance

Résumé

Contexte. – La surveillance du *Staphylococcus aureus* résistant à la méticilline (SARM) est une priorité nationale. La mesure de son taux représente l'un des six indicateurs du tableau de bord piloté par le ministère de la santé. Depuis 2002, l'Institut de veille sanitaire (InVS) réalise une surveillance annuelle des SARM qui permet d'estimer une densité d'incidence. Aujourd'hui, le Programme de médicalisation des systèmes d'information (PMSI) pourrait-il être utilisé pour cette surveillance ? Il s'agit de proposer une méthode permettant au niveau national et à partir des données issues du PMSI de décrire les densités d'incidences des SARM et d'analyser les écarts avec les résultats publiés par l'InVS pris comme références.

Méthodes. – Les données de la base nationale PMSI pour les années 2006 à 2009 ont été utilisées. Les données de références étaient celles publiées par l'InVS de 2006 à 2009. Une densité de SARM correspondant au nombre de SARM codé dans l'année rapporté à 1000 journées

* Auteur correspondant.
Adresse e-mail : catherine.quantin@chu-dijon.fr (C. Quantin).

0398-7620/$ – see front matter
http://dx.doi.org/10.1016/j.respe.2013.04.008

d'hospitalisations a été calculé pour la comparaison avec les densités d'incidence mesurées par l'InVS. L'évolution du recueil des SARM dans le PMSI a été modélisée par une régression de Poisson.

Résultats. – La densité d'incidence mesurée par l'InVS est plus élevée que la densité de SARM calculée à partir du PMSI mais cet écart semble s'atténuer au cours du temps. Le rapport densité PMSI/densité d'incidence InVS des SARM était de 0,8 % en 2006 et environ 9,2 % en 2009. Nous avons observé des tendances inversées avec une évolution croissante de la densité de SARM identifié par le PMSI. De plus l'année était associée significativement à la densité d'incidence ($p = 0{,}01$).

Conclusion. – La base PMSI représente une source complémentaire de données pour la surveillance des SARM en milieu hospitalier, notamment en apportant ses capacités de dédoublonnage régionale et nationale grâce au chaînage et d'estimer des densités d'incidence dans des établissements non participants à la surveillance.

Mots clés : *Staphylococcus aureus* résistant à la méticilline (SARM) ; Infection croisée ; Programme de médicalisation du système d'information (PMSI) ; Régression ; Surveillance

1. Introduction

La problématique liée au *Staphylococcus aureus* résistant à la méticilline (SARM) est liée au fait qu'il n'est pas uniquement une des causes fréquentes d'infections nosocomiales, mais également à l'origine d'infections communautaires parfois sévères pour certaines souches [1]. Les SARM font partie des bactéries multirésistantes dont la surveillance au sein des établissements de santé est une priorité nationale.

Il existe au niveau national un indicateur de SARM publié depuis 2007 et mis à la disposition uniquement de l'ensemble des établissements de santé (ceux réalisant au moins 30 000 journées d'hospitalisations par an). Il s'agit d'un classement de performance qui repose sur les classes de percentiles (10, 30, 70, 90) calculées à partir du taux triennal de SARM (une agrégation sur trois ans du nombre de patients hospitalisés chez lesquels au moins une souche de SARM a été isolée dans l'année au sein d'un prélèvement à visée diagnostique rapporté à 1000 journées d'hospitalisations sans distinction entre les cas acquis et ceux importés) [2]. Tous les établissements sont donc tenus de fournir les données nécessaires au calcul de cet indicateur dont l'objectif est de permettre de déduire pour chacun des tendances évolutives des taux de SARM. Parallèlement, l'Institut de veille sanitaire (InVS) publie chaque année depuis 2006 un autre indicateur de SARM accessible à tous : la densité d'incidence de SARM qui est le nombre absolu de patients porteurs de BMR acquis et détecté sur la base des prélèvements à visée diagnostique positifs rapportés à 1000 journées d'hospitalisations. L'objectif ici étant la maîtrise de la diffusion des SARM. Les données utilisées pour le calcul de cet indicateur résultent des enquêtes de surveillance encore appelées réseau Raisin (Réseau d'alerte, d'investigation et de surveillance des infections nosocomiales), réalisées au sein des établissements volontaires, chaque année durant le premier semestre et pendant trois mois consécutifs. Les résultats publiés sur un rythme annuel sont d'abords analysés au niveau de chaque centre de coordination de la lutte contre les infections nosocomiales (CClin) et ensuite agrégés au niveau national.

Depuis le 1er janvier 2011, les établissements doivent mettre à la disposition du public les résultats publiés chaque année des indicateurs de qualité et de sécurité des soins parmi lesquels figure l'indicateur de SARM (ou classe de performance) [3]. Il s'agissait du troisième indicateur figurant dans le tableau de bord des infections nosocomiales piloté par le ministère de la Santé [2], qui comportait l'indicateur composite des activités de lutte contre les infections nosocomiales (Icalin) et l'indicateur de volume de produits hydro-alcooliques (SHA) consommés.

Le Programme de médicalisation du système d'information (PMSI), dont l'objectif principal est la tarification à l'activité des établissements de santé, est aujourd'hui généralisé à l'ensemble des établissements ayant une activité de court séjour de médecine, de chirurgie et/ou d'obstétrique (MCO). Ce système permet un recueil, standardisé, des informations médico-administratives sur les séjours hospitaliers des patients directement sur site. Ces informations sont sauvegardées dans une base de données nationale.

Il s'agissait dans ce travail d'analyser dans quelle mesure le PMSI pouvait être associé ou intégré dans l'arsenal de la surveillance des SARM comme outil de recueil des informations médicales utiles à l'identification de ces derniers. Cela permettrait ainsi à chaque établissement de pouvoir construire l'un ou l'autre des indicateurs des SARM selon la méthodologie recommandée.

L'objectif de ce travail était donc de réaliser un état des lieux du recueil des SARM dans le PMSI depuis 2006 jusqu'en 2009 et d'évaluer l'intérêt de cette base comme outil complémentaire pouvant contribuer à la surveillance nationale des SARM. Cette évaluation a été effectuée en comparant les mesures de densité d'incidence et les densités de SARM proposée par l'InVS et publiées annuellement dans le cadre des résultats de la surveillance Raisin depuis 2006 jusqu'en 2009. Il s'agissait également d'analyser l'évolution de ce recueil dans cette même période.

2. Population et méthodes

2.1. Les données

Ce travail a nécessité l'exploitation de trois sources de données différentes à savoir les bases nationales PMSI, les rapports annuels sur la surveillance des bactéries multirésistantes (BMR) publiés par l'InVS et les données de la statistique annuelle des établissements.

Tout d'abord, les bases nationales PMSI (dans la suite nous parlerons de bases PMSI) pour les années 2006 à 2009 ont été

utilisées. Elles contenaient les informations médico-administratives de l'ensemble des séjours hospitaliers qui se sont déroulés pendant cette période sur la France entière, quel que soit l'établissement de santé public ou privé. Après autorisation de la Commission nationale de l'informatique et des libertés (CNIL – N° Autorisation 1332655), ces bases nous ont été fournies par l'Agence technique de l'information sur l'hospitalisation (ATIH).

Ensuite, nous nous sommes intéressés aux résultats publiés dans les rapports annuels (2006 à 2009) de « Surveillance des bactéries multirésistantes dans les établissements de santé en France » [4–7]. Nous avons limité notre champ à la densité d'incidence des SARM, mesurée dans les services de médecine, de chirurgie et d'obstétrique. Les résultats concernant la psychiatrie ou les soins de suites n'ont pas été pris en compte.

Et enfin, pour identifier et caractériser les établissements de santé, nous nous sommes servis des données du Fichier national des établissements sanitaires et sociaux (FINESS) et des données de la Statistique annuelle des établissements (SAE), dans leurs versions disponibles en ligne [8,9].

2.2. Sélection des séjours

Les bases PMSI contiennent des informations médico-administratives synthétiques et standardisées concernant les séjours. Les informations médicales, en l'occurrence les diagnostics, figurent sous forme codée selon la plus récente version de la Classification internationale des maladies (CIM 10) de l'Organisation mondiale de la santé (OMS), complétée, le cas échéant, d'extensions publiées par l'ATIH.

Le codage de l'information des SARM fait appel à deux codes devant être présents simultanément : d'une part, celui de l'agent infectieux le *Staphylococcus aureus* (B95.6) et, d'autre part, le code de la résistance à la méticilline (U80.1).

Les séjours d'hospitalisation ayant une durée totale supérieure à un jour et présentant le codage du SARM étaient retenus. Ce codage était recherché uniquement en diagnostic associé significatif (DAS). En effet, dans le résumé du PMSI, la morbidité est codée sous une forme hiérarchique : d'abord le diagnostic principal (DP) qui correspond au problème de santé ayant motivé l'admission du patient, puis le diagnostic relié (DR) recueilli dans certaines situations et qui apporte une précision sur le DP. Enfin, un (ou plusieurs) DAS désigne un problème de santé pris en charge en plus du DP. Ainsi, la codification des SARM doit être saisie comme DAS uniquement pour être valide. Une saisie par exemple du code B95.6 ou U80.1 comme un DP est une erreur de codage.

Lorsqu'un séjour était identifié, un certain nombre d'informations étaient retenues, dont l'année du séjour, sa durée, la notion « d'isolement (prophylactique) » du patient indiquée par un code diagnostique (Z29.0) et l'identifiant unique des entités juridiques des établissements. En effet, cette information était croisée avec le fichier Finess et les données de SAE de manière à pouvoir décrire les établissements de santé en termes de catégorie d'établissement (centre hospitalier universitaire [CHU], centre hospitalier [CH], clinique privée [CL], centre de lutte contre le cancer [CLCC], etc.), de nombre de lits déclarés ou encore de Centre de coordination et de lutte contre les infections nosocomiales (CClin) d'appartenance.

2.3. Méthodes statistiques

L'unité statistique était le séjour du patient.

Pour comparer les densités de SARM calculées à partir de la base PMSI avec les densités d'incidences de SARM publiées par l'InVS, nous avons pris en compte uniquement les séjours de plus de 24 heures. Nous avons travaillé avec les mêmes catégories d'établissement que celles utilisée par l'InVS [4–7]. La méthode de sélection des bases PMSI réalisée ne permettant pas d'estimer selon des critères comparables au protocole Raisin le caractère incident ou non de l'infection à SARM diagnostiqué, une densité de SARM a été calculé à partir des bases PMSI. La densité de SARM correspond au nombre de SARM codés dans l'année rapporté à 1000 journées d'hospitalisations. Du point de vue du PMSI, nous nous sommes intéressés à l'activité de court séjour du champ MCO de ces établissements. En revanche, il ne nous a pas été possible de repérer de manière spécifique les SARM codés par type de spécialités médicales des services. En effet, dans la base nationale PMSI, les résumés d'un séjour multi-services dans l'établissement comporte l'ensemble des diagnostics associés significatifs (DAS) notés pendant le séjour, sans qu'il soit possible d'identifier le service ayant codé l'un ou l'autre des DAS. Le champ du MCO est celui qui présente aujourd'hui une bonne exhaustivité du recueil. Les catégories d'établissements étudiées étaient les suivantes : les CHU, les CH qui regroupaient les ex-hôpitaux locaux, les CL incluaient les établissements privés participant aux services publics, les CLCC et les services de santé des armées (SSA). Les établissements ont été regroupés selon les cinq CClin existant : CClin-Est, CClin-Ouest, CClin-Sud-Est, CClin-Sud-Ouest, CClin-Paris-Nord hors Assistance publique–Hôpitaux de Paris (AP–HP) et CClin-Paris-Nord AP–HP.

Les intervalles de confiances pour les proportions ont été calculés avec la méthode binomiale exacte. L'étude de l'évolution du recueil a été réalisée avec une modélisation de la densité de SARM dans le PMSI, à partir d'une régression de Poisson et en utilisant les covariables suivantes : l'année, le CClin, le nombre d'établissements et le nombre de lits par CClin. L'ensemble des analyses a été réalisé à l'aide du logiciel SAS 9.3.

3. Résultats

Les bases PMSI de 2006 à 2009 contenaient pour ces quatre années un nombre moyen de 22 795 915 séjours avec une valeur minimum de 21 201 102 pour l'année 2007 et une valeur maximum de 24 575 239 en 2009. De même, pour ce qui concernait les patients, on observait une moyenne de 11 637 430 patients avec une augmentation régulière au fil du temps (de 10 697 203 en 2006 à 12 691 485 en 2009) (Tableau 1).

Les entités juridiques des établissements de santé ont été regroupées selon leur appartenance à un CClin. Le nombre

Tableau 1
Répartition du nombre de patients et de séjours par années dans les bases nationales du PMSI–MCO.

Année	Nombre de séjours	Nombre de patients
2006	21 578 428	10 697 203
2007	21 201 102	11 047 482
2008	23 828 892	12 113 551
2009	24 575 239	12 691 485

PMSI : programme de médicalisation du système d'information ; MCO ; médecine, de chirurgie et/ou d'obstétrique.

minimum d'entités juridiques [1] était observé au niveau du CClin Paris-Nord AP–HP et le nombre maximum moyen (381 ± 7) était observé au niveau du CClin Sud-Est. Le Tableau 2 présente une répartition sur quatre années (2006 – 2007 – 2008 – 2009) du nombre moyen d'entités juridiques par CClin et par catégories d'établissement calculé à partir des données du PMSI. Le nombre d'établissement privés (726) est supérieur au nombre cumulé de CHU (32) et de CH (637).

Le nombre total de SARM identifiés par année dans la base PMSI était variable de 2006 à 2009. Le minimum était observé en 2006 (226) et le maximum en 2009 (2349). Cette tendance croissante était retrouvée au niveau des catégories des établissements comme le montre le Tableau 3. En effet, le nombre absolu de SARM identifié par exemple dans les CH était de 120 en 2006, ensuite de 202 (+68 %) en 2007, puis 310 (+53 %) en 2008 et enfin, 1282 (+314 %) en 2009. Nous observons également dans ce même tableau les nombres absolus des SARM et des isolements (prophylactiques) codés pendant le même séjour. La seconde valeur étant toujours inférieure à la première et cela quel que soit le type d'établissement.

Les résultats du calcul de la densité des SARM à partir des bases PMSI et les valeurs extraites des publications de l'InVS sont résumés dans le Tableau 4 sous la forme d'un rapport densité PMSI/densité d'incidence InVS des SARM. Nous avons observé et cela pour toutes les catégories d'établissements, que ce rapport évoluait de 0,8 % en 2006 jusqu'à 9,2 % en 2009. Au niveau des CH spécifiquement ce chiffre était passé de 1 % en 2006 à 12 % en 2009, soit une tendance à la baisse de l'écart du nombre de SARM détecté entre le PMSI et l'InVS.

Tableau 2
Répartition du nombre moyen des entités juridiques sur quatre années (2006–2007–2008–2009) par catégories d'établissements et par CClin à partir des données du PMSI (base nationale 2006–2009).

CClin/catégorie établissement	CH	CHU	CL	CLCC	SSA	Total
CClin/Est	99	7	80	4	0	190
CClin/Ouest	128	7	97	4	0	236
CClin/Paris-Nord AP–HP	0	1	0	0	0	1
CClin/Paris-Nord hors AP–HP	123	3	234	5	2	368
CClin/Sud-Est	174	8	194	5	0	381
CClin/Sud-Ouest	113	6	121	2	0	242
Total	637	32	726	20	2	1417

CHU : centre hospitalier universitaire ; CH : centre hospitalier ; CL : clinique privées ; CLCC : centre de lutte contre le cancer ; SSA : service de santé des armées ; CClin : centre de coordination et de lutte contre les infections nosocomiales ; PMSI : programme de médicalisation du système d'information.

Cette évolution dans les sens opposés est observable au niveau de la Fig. 1 : l'image (a) montre la décroissance de l'indicateur de SARM telle que mesurée par l'InVS au travers d'enquêtes d'incidences, l'image (b) montre « la montée en charge » du recueil de SARM dans le PMSI et l'image (c) est une superposition des deux droites de tendances précédentes.

Les résultats de la modélisation sont présentés dans le Tableau 5. Seule la variable année était statistiquement liée à la densité d'incidence des SARM mesurée à partir des données du PMSI ($p = 0{,}01$). De plus, la valeur négative des coefficients (la modalité de référence était l'année 2009) semblait cohérente avec la tendance croissante (au fil des années) observée sur la Fig. 1.

Si nous posons l'hypothèse que les tendances aussi bien au niveau des données de l'InVS que celles du PMSI se maintiennent dans les années à venir, nous pouvons déterminer que les deux droites (Fig. 1c) pourraient se croiser entre 2013 et 2014 avec les valeurs de densité de SARM variant entre 0,29 et 0,38.

4. Discussion

Nous avons observé que le nombre de cas de SARM détecté était toujours en progression chaque année depuis 2006 mais encore en dessous des chiffres publiés par l'InVS qui sont considérés comme une référence nationale. L'évolution du rapport densité PMSI/densité d'incidence InVS des SARM toutes catégories d'établissements était de 0,8 % en 2006, ensuite de 1,2 % en 2007 puis de 1,9 % en 2008 et enfin de près de 9,2 % en 2009. Ces chiffres sont encourageants au regard de la difficulté qui existe dans le recueil de ces informations en routine. Nous avons également montré que cette progression annuelle était statistiquement significative ($p = 0{,}01$). Lorsque l'on superpose les droites de tendances des densités pour l'InVS et le PMSI, et avec l'hypothèse du maintien de ces deux tendances (décroissante pour l'InVS et croissante pour le PMSI) dans les prochaines années, on peut déterminer un croisement de droite entre 2013 et 2014 avec des valeurs de densité entre 0,29 et 0,38.

Les résultats présentés dans cet article montrent que l'utilisation du PMSI comme un outil complémentaire de la surveillance Raisin dans la surveillance des SARM semble possible et permet d'apporter une réponse à la question posée de son utilisation dans la surveillance des infections nosocomiales [10]. Toutefois, il est important de pouvoir garantir une bonne exhaustivité du recueil des informations nécessaires de façon à réduire les erreurs liées à une utilisation trop précoce. En effet, dès 2006 l'identification des SARM dans le PMSI est possible. Mais les résultats du calcul du nombre de SARM rapporté à 1000 journées d'hospitalisations de 2006 à 2009 y montrent une sous-estimation importante par rapport aux résultats de la surveillance Raisin publiés par l'InVS. L'écart était en 2006 d'un SARM identifié dans le PMSI pour 125 identifiés dans la surveillance Raisin toutes catégories d'établissements confondues et en 2009 il n'était plus que d'environ 9,2 %. La confrontation de ces deux sources de données nous permet de considérer que depuis 2006, les efforts

Tableau 3
Répartition du nombre de SARM et celui correspondant aux d'isolements prophylactiques par catégories d'établissements et par année à partir des données du PMSI (base nationale 2006–2009).

Catégories d'établissements	2006		2007		2008		2009	
	SARM	Z29.0	SARM	Z29.0	SARM	Z29.0	SARM	Z29.0
CH	120	17	202	51	310	73	1282	284
CHU	27	3	86	12	125	17	687	275
CL	79	18	90	26	94	19	360	127
CLCC	0	0	0	0	0	0	20	8
Total	226	38	378	89	529	109	2349	694

SARM : *Staphylococcus aureus* résistant à la méticilline ; CHU : centre hospitalier universitaire ; CH : centre hospitalier ; CL : clinique privées ; CLCC : centre de lutte contre le cancer ; SSA : service de santé des armées ; PMSI : programme de médicalisation du système d'information ; Z29.0 : code C1M10 de l'isolement (prophylactique).

fournis pour l'amélioration de l'exhaustivité du recueil des SARM dans les bases PMSI sont plutôt encourageants dans la mesure où nous avons pu constater une évolution annuelle significative ($p = 0{,}01$). De plus, compte tenu des enjeux budgétaires du PMSI au sein des établissements de santé, l'hypothèse de la poursuite de l'amélioration de la qualité du recueil est défendable.

La possibilité d'identifier dans le PMSI les séjours de patients avec un prélèvement positif à SARM n'est pas suffisante pour garantir une exhaustivité complète des cas. De même que les résultats de la surveillance Raisin ne sont pas facilement extrapolables au niveau national. Mais le PMSI possède des atouts très intéressants en tant que source de données pour des études de portée nationale. Il représente une source d'information très intéressante du fait de son mode de recueil qui est à la fois généralisé, régulier et standardisé. En ce sens, certains auteurs s'accordent à penser que ce mode de recueil se rapproche de celui d'une étude de cohorte dans laquelle tous les centres d'inclusions de patients (même les futurs) seraient déjà favorables et dont les thèmes d'études resteraient à définir, et qu'il serait suffisamment mature pour être utilisé dans les études épidémiologiques longitudinales [11]. En effet, la possibilité du chaînage des séjours d'un patient dans le respect de la confidentialité des informations médicales grâce à l'utilisation du numéro MAGIC anonyme chaînable est une avancée primordiale pour les travaux épidémiologiques français. Ainsi on retrouve dans la littérature, des études concernant les SARM qui ont utilisé des bases médico-administratives comme source de données [12,13] mais pas à une telle échelle nationale.

Tableau 4
Les rapports densités PMSI/densités d'incidences InVS des SARM pour 1000 journées d'hospitalisations calculées par catégories d'établissements et par année.

Catégorie d'établissement	2006	2007	2008	2009	Total
CH	0,010	0,016	0,031	0,116	0,041
CHU	0,003	0,005	0,013	0,107	0,023
CL	0,013	0,015	0,011	0,049	0,024
CLCC	0,000	0,000	0,000	0,100	0,019
SSA	0,000	0,000	0,000	0,000	0,000
Total	0,008	0,012	0,019	0,092	0,030

PMSI : programme de médicalisation du système d'information ; InVS : institut national de veille sanitaire ; SARM : *Staphylococcus aureus* résistant à la méticilline ; CHU : centre hospitalier universitaire ; CH : centre hospitalier ; CL : clinique privées ; CLCC : centre de lutte contre le cancer ; SSA : service de santé des armées.

L'intérêt de ce travail était d'envisager dans quelle mesure le PMSI permettrait une amélioration de la surveillance des SARM. Du fait de la portée nationale du recueil, et des possibilités de chaînage tout en respectant la confidentialité des informations médicales nous avons ainsi la possibilité de tracer le parcours d'un patient qui serait porteur d'un SARM. Ce chaînage réduirait ainsi le risque d'une redondance dans le dénombrement dans le cas d'un transfert d'un établissement à un autre, contrairement aux enquêtes réalisées par l'InVS, qui se font indépendamment dans chacun des établissements participants.

Ce travail présente quelques limites qui sont importantes à souligner. Les résultats du calcul du nombre de SARM isolé d'un prélèvement à visée diagnostique et rapporté à 1000 journées d'hospitalisations, réalisés à partir de deux sources de données distinctes (InVS et PMSI) ont été présentés dans les limites des capacités communes aux deux sources. Contrairement à la surveillance Raisin, il n'était pas possible au niveau du PMSI de tracer l'information du SARM jusqu'au niveau fin de la spécialité médicale du service (réanimation par exemple) ayant hébergé le patient pendant son séjour. Cela tient à la structure des résumés de séjours anonymes (RSA) enregistrées dans le PMSI qui désolidarisent les DAS du service producteur de l'information lorsque le patient a fréquenté plusieurs services pendant le même séjour. De même tout association du SARM avec un code diagnostique d'infection ou d'isolement (prophylactique) ne reflète pas totalement la réalité de la pathologie prise en charge et cela pour la raison indiquée plus haut. Un recueil aussi fin des informations ne pourrait à ce jour malheureusement se concevoir qu'au niveau local d'un établissement. Une autre limite tient du fait que le PMSI est un outil dont la finalité est la tarification à l'activité des établissements de santé. Il n'est donc pas considéré comme un instrument pouvant servir l'épidémiologie. Pourtant cette base de données représente une possibilité de rassembler des informations d'origines diverses participant à la prise en charge du patient. Ce sont par exemples les informations de bactériologie (germes et résistances), de la pharmacie (traitements) ou encore des services cliniques. De plus, le fait que son alimentation au quotidien par les professionnels de

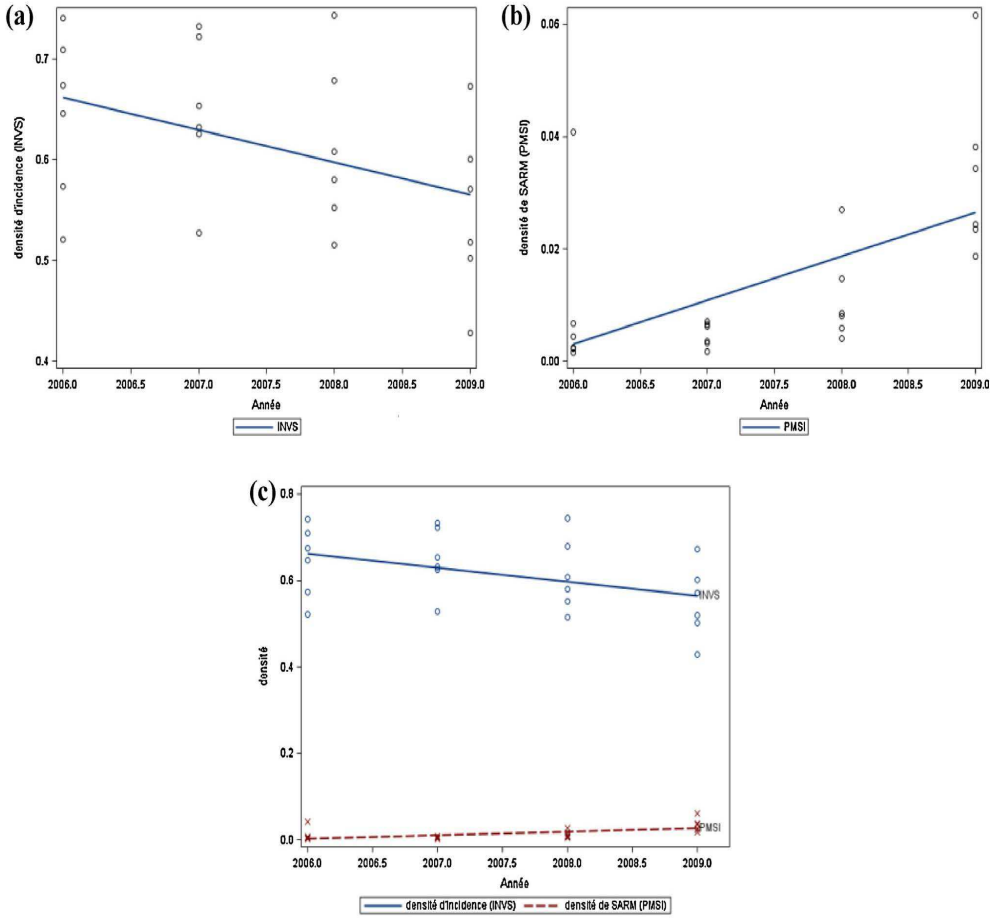


Fig. 1. Évolution annuelle entre 2006 et 2009 de la densité d'incidence des *Staphylococcus aureus* résistant à la méticilline (SARM) calculé par l'Institut nationale de veille sanitaire (InVS) (a) et de la densité de SARM à partir de la base Programme de médicalisation du système d'information (PMSI) (b) : a : droite de tendance pour les données de référence de l'InVS ; b : droite de tendance pour les données issues du PMSI ; c : superposition des droites de tendances (a) et (b).

santé soit parfois vécue comme une contrainte pénaliserait sa qualité et son exhaustivité. Le contraste assez surprenant entre les données de l'InVS et celles du PMSI trouve une explication dans le fait que ces données sont recueillies dans des contextes différents. En effet, les données de l'InVS sont recueillies, au sein des établissements, dans le cadre d'une enquête nationale de surveillance des bactéries multi-résistantes par les équipes opérationnelles d'hygiène (EOH). Cette enquête se déroule sur une période certes plus courte (trois mois), mais son objectif unique conduit probablement à une meilleure « précision » du recueil. De leur côté, les données du PMSI sont recueillies pour une grande partie dans les services cliniques et/ou au sein des départements de l'information médicale (DIM), selon que l'établissement réalise un codage centralisé ou décentralisé des séjours. Cette activité, qui s'intègre, dans le cadre du recueil informatisé des données médicales du séjour d'un patient, a été en quelque sorte ajoutée aux nombreuses autres activités des personnels des établissements et est souvent vécue comme contraignante. Certains recueils vont donc à l'essentiel, c'est-à-dire enregistrent les éléments directement requis pour la valorisation des séjours et l'information sur l'identification d'un SARM est négligée.

Tableau 5
Résultats de la modélisation de la densité de SARM.

Paramètres	Coefficients	IC 95	*p*
Année référence : 2009			
2006	−1,2	[−2,16 ; −0,25]	0,01
2007	−1,36	[−2,38 ; −0,35]	0,01
2008	−0,98	[−1,71 ; −0,25]	0,01
CClin référence : Ouest			
Est	2,1	[−1,26 ; 5,45]	0,22
Paris-Nord AP–HP	12,15	[−5,68 ; 29,98]	0,18
Paris-Nord hors AP–HP	−6,46	[−15,63 ; 0,17]	0,17
Sud-Est	−6,6	[−20,61 ; 7,41]	0,36
Sud-Ouest	−0,82	[−4,79 ; 3,14]	0,68
Établissements (nombre)	0,06	[−0,01 ; 0,13]	0,08
Lits (nombre)	0		0,91

Des mesures visant à pallier ces difficultés sont déjà entreprises dans les établissements. Pour aller plus loin, il conviendrait dans la mesure du possible de synchroniser les

activités de recueil des données de l'InVS et du PMSI. Cela pourrait passer par une reconnaissance au niveau national de l'importance de cette problématique, grâce non seulement à une valorisation du travail des EOH dans la surveillance des SARM, mais également par une reconnaissance de l'importance du codage des résistances à la méticilline (souvent associées à la présence d'un *Staphylococcus aureus*). Ainsi, la reconnaissance de ce code comme un diagnostic associé significatif pendant le séjour, par l'attribution d'un niveau de sévérité d'indice 3 en 2010 puis de 4 en 2011 (sur une échelle de 1 à 4), traduit déjà le fait que la présence de SARM conduit à un allongement de la durée de séjour, puisque les minima vont de 0 jour (niveau 1) à quatre jours (niveau 4). Cette évolution du niveau de sévérité permet donc au recueil des SARM d'influer directement sur la valorisation des séjours.

Cet encouragement au codage par sa valorisation tarifaire permet ainsi d'améliorer l'exhaustivité du recueil et d'envisager d'étendre l'utilisation du PMSI au niveau national. En effet, si le recueil des données relatives aux SARM était exhaustives au niveau national, la reconstitution d'un parcours intra- et/ou extrahospitalier d'un patient porteur de SARM et identifié comme tel serait alors possible, grâce au chaînage des séjours d'un patient, entraînant, par voie de conséquence, la possibilité de dédoublonnage des patients transférés d'un établissement à un autre. Il est très important de pouvoir nuancer cette forme d'encouragement au recueil des SARM qui serait basée uniquement sur des décisions tarifaires car cela pourrait à terme nuire à la qualité de la surveillance épidémiologique des SARM.

Une autre forme d'encouragement pourrait être de réduire le nombre de codes à saisir en créant par exemples de nouveaux codes spécifiques pour des infections à SARM.

5. Conclusion

Le recueil des SARM dans le PMSI est en progression, pour l'instant ce dernier ne remplacera pas une surveillance, cette base PMSI peut venir en appoint lors des enquêtes de l'InVS si et seulement si la tendance à la hausse du recueil se confirme dans les prochaines années. L'indicateur calculé par l'InVS reste la référence pour suivre la diffusion des SARM au niveau national. À une échelle locale (d'établissement), l'utilisation des données du PMSI permettrait de disposer rapidement des résultats de surveillance ainsi que des tendances et surtout de proposer très précocement des mesures d'ajustements.

Déclaration d'intérêts

Les auteurs déclarent ne pas avoir de conflits d'intérêts en relation avec cet article.

Références

[1] Tattevin P. Les infections à *Staphylococcus aureus* résistant à la meticilline (SARM) d'acquisition communautaire. Med Mal Infect 2011;41:167–75.

[2] Circulaire N° DHOS/ED/DGS/SD5C/2006/163 du 7 avril 2006 relative au tableau de bord des infections nosocomiales et portant sur les modalités de calcul de l'indicateur sur le taux de *Staphylococcus aureus* résistant à la méticilline par les établissement de santé. 2006.

[3] Arrêté du 28 décembre 2010 fixant les conditions dans lesquelles l'établissement de santé met à la disposition du public les résultats publiés chaque année des indicateurs de qualité et de sécurité des soins. 2010.

[4] Jarlier V, Arnaud A, Carbonne A. Surveillance des bactéries multirésistantes dans les établissements de santé en France. Réseau BMR-Raisin – Résultats 2006. 2009.

[5] InVS-RAISIN. Surveillance des bactéries multirésistantes dans les établissements de santé en France. Réseau BMR-Raisin – Résultats 2007. 2009.

[6] Jarlier V, Arnaud I, Carbonne A. Surveillance des bactéries multirésistantes dans les établissements de santé en France. Réseau BMR-Raisin – Résultats 2008. Saint-Maurice: Institut de veille sanitaire; 2010.

[7] Jarlier V, Arnaud I, Carbonne A. Surveillance des bactéries multirésistantes dans les établissements de santé en France. Réseau BMR-Raisin – Résultats 2009. 2011.

[8] Statistiques annuelles des établissements de santé (SAE) [En ligne] [Internet]. 2008. Available from: http://www.sae-diffusion.sante.gouv.fr.

[9] DREES. Le Fichier Nationale des Établissements Sanitaires et Sociaux (FINESS) [en ligne] 2010. Available from: http://finess.sante.gouv.fr/jsp/index.jsp.

[10] Gerbier S, Bouzbid S, Pradat E, Baulieux J, Lepape A, Berland M, et al. Intérêt de l'utilisation des données du Programme médicalisé des systèmes d'information (PMSI) pour la surveillance des infections nosocomiales aux Hospices Civils de Lyon. Rev Epidemiol Sante Publique 2011;59:3–14.

[11] Olive F, Gomez F, Schott AM, Remontet L, Bossard N, Mitton N, et al. Analyse critique des donnees du PMSI pour l'epidemiologie des cancers : une approche longitudinale devient possible. Rev Epidemiol Sante Publique 2011;59:53–61.

[12] Edris B, Eid S, Molitoris A, Reed JF. Incidence and potential financial impact of resistant *Staphylococcus aureus* in an Academic Community Hospital. Internet J Infect Dis 2008;2:16.

[13] Fourquet F, Demont F, Lecuyer AI, Rogers MA, Bloc DH. PMSI et surveillance des infections nosocomiales : théorie et faisabilité [French medical hospital information system and cross infection surveillance: theory and feasibility]. Med Mal Infect 2003;33:110–3.

CONCLUSION-DISCUSSION

Sur les deux questions traitées, le point commun qui a fortement orienté notre travail était l'aspect évolutif anticipé dans le recueil des données. Le changement était temporel pour les 2 bases de données : la description dans la base médico-administrative de plusieurs séjours hospitaliers différents pour un même patient, et pour l'étude clinique, l'administration d'un auto-questionnaire de qualité de vie à des temps différents et prédéterminés dans le protocole. Un changement spatial était présent dans la base médico-administrative dans la mesure où les différents séjours des patients ne se déroulaient pas toujours ni dans le même établissement, ni dans le même département et ni dans la même région. C'est donc fort de cet aspect longitudinal des données que nous avons conduit l'exploration. Nous avons donc pu construire pour chaque patient soit une trajectoire hospitalière de prise en charge soit un profil individuel de la qualité de vie. L'hétérogénéité importante observée chez les patients, sur la description des trajectoires hospitalières comme au niveau des profils individuels de qualité de vie nous a ainsi conduits à construire une typologie des trajectoires et celle des profils. Une description des composantes de la typologie à partir des informations liées aux patients nous a permis d'avancer dans l'analyse des problèmes posés. La représentation graphique adoptée dans les deux cas à savoir les barres d'histogramme était une nouveauté dans le domaine médical.

L'intérêt majeur de ces méthodologies réside dans le fait qu'elles apportent des aiguillages pertinents sur les questions posées. Ainsi grâce à l'analyse spatiale nous avons pu constater et estimer les « fuites régionales » dans la prise en charge chirurgicale dans les cancers du poumon chez les patients bourguignons. De même la mise en évidence des 3 composantes de la typologie des profils évolutifs de QdV montre bien qu'il est possible de retrouver par l'analyse des données une information apparemment intuitive. La contrepartie est une mise en œuvre pratique des éléments techniques qui nécessite une certaines compétences d'une part dans l'analyse structurelle des données mais aussi dans la manipulation des nouveaux outils statistiques (1-3). Une bonne compréhension de la structure des données à analyser peut conduire à oser des stratégies qu'il n'était possible d'envisager avant à savoir l'imputation des données avec comme seul objectif, d'évaluation statistiques. La contrainte est bien entendu de pouvoir au terme de l'analyse isoler l'impact dû à l'imputation des données dans les résultats obtenus d'où l'analyse de sensibilité réalisées dans la construction des profils évolutifs de qualité de vie. Nous n'avons pas retrouvé dans la littérature de travaux similaires mais les résultats satisfaisants de certaines techniques nous ont permis d'argumenter certains de nos choix (4-8).

Dans les deux d'études qui ont été décrits, la méthodologie ne permettait pas d'obtenir des certitudes par rapport aux questions posées mais plutôt de dégager différentes pistes de réflexions. Avec l'application de ces stratégies, à défaut d'obtenir des certitudes, elles permettent en quelque chose de ciblée

Ce travail montre l'intérêt de l'utilisation des méthodes de fouilles de données pour l'analyse des trajectoires dans le domaine de la santé à partir de deux exemples de problématique (l'un relatif aux trajectoires de soins, et l'autre relatif aux trajectoires de qualité de vie) et de deux types de bases de données (données PMSI et données d'essais thérapeutiques).

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Billard L, Diday E. Symbolic Data Analysis: Conceptual Statistics and Data Mining: WILEY; 2006.

2. Diday E. Introduction a l'analyse des donnees symboliques. 1989:38.

3. DIDAY E, NOIRHOMME-FRAITURE M. *Symbolic Data Analysis and the SODAS Software*. Wiley, editor2008.

4. Quantin C, Billard L, Touati M, Andreu N, Afonso F, Battaglia G, et al. Classification and Regression Trees on Aggregate Data Modeling: An Application in Acute Myocardial Infarction. Journal of Probability and Statistics. 2011.

5. Leffondre K, Abrahamowicz M, Regeasse A, Hawker G, Badley E, McCusker Jea. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. J Clin Epidemiol. 2004;57(10):1049-62.

6. Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J. A Framework for Multiple Imputation in Cluster Analysis. Am J Epidemiol2013.

7. Breaban M, Luchian H. A unifying criterion for unsupervised clustering and feature selection. Pattern Recognition. 2011;44(4):854-65.

8. Edwin D. Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. Revue de Statistique appliquée. 1971;19(2):19-33.

Résumé

Contexte

Dans le domaine de la santé, l'analyse des données pour l'extraction des connaissances est un enjeu en pleine expansion. Les questions sur l'organisation des soins ou encore l'étude de l'association entre le traitement et qualité de vie (QdV) perçue pourraient être abordées sous cet angle. L'évolution des technologies permet de disposer d'outils de fouille de données performants et d'outils statistiques enrichis de méthode avancées, utilisables par les non-experts. Nous avons illustré cette méthode au travers de deux questions d'actualité :1 / Quelle organisation des soins pour la prise en charge des cancers ? 2/ étude de la relation chez les patients souffrant d'un cancer métastatique entre la QdV liée à la santé perçue et les traitements reçus dans le cadre d'un essai thérapeutique.

Matériels et méthodes

Nous disposons aujourd'hui de volumineuses bases de données. Certaines retracent le parcours hospitalier des patients, comme c'est le cas pour les données d'activités hospitalières recueillies dans le cadre du programme de médicalisation des systèmes d'information (PMSI). D'autres conservent les informations sur la QdV perçues par les patients et qui recueillies en routine actuellement dans les essais thérapeutiques. L'analyse de ces données a été réalisée suivant trois étapes principales : Tout d'abord une étape de préparation des données dont l'objectif était la compatibilité à un concept d'analyse précisé. Il s'agissait par exemple de transformer une base de données classique (centrée sur le patient) vers une nouvelle base de données où « l'unité de recueil » est une entité autre que le patient (ex. trajectoire de soins). Ensuite une deuxième étape consacrée à l'application de méthodes de fouille de données pour l'extraction connaissances : les méthodes d'analyse formelle des concepts ou encore les méthodes de classifications non-supervisée. Et enfin l'étape de restitution des résultats obtenus et présenté sous forme graphique.

Résultats

Pour la question de l'organisation des soins, nous avons construit une typologie des trajectoires hospitalières des soins permettait de réaliser un état des lieux des pratiques dans la prise en charge des cancers étudié depuis la chirurgie jusqu'à un an de suivi des patients. Dans le cas du Cancer du sein, nous avons décrit une typologie de prise en charge sur la base des coûts d'hospitalisation sur un suivi d'un an. Pour la deuxième question, nous avons également construit une typologie des profils évolutifs de la QdV. Celle-ci comportait 3 classes : une classe d'amélioration, une classe de stabilité et une classe de dégradation.

Conclusion

L'intérêt majeur de ce travail était de mettre en évidence des pistes de réflexion permettant des avancées dans la compréhension et la construction de solutions adaptées aux problèmes.

Mots clés

Fouille de données ; Classification ; Cancers ; trajectoire de soins ; Qualité de vies ; Imputation de données.