Cancer patient flows discovery in DRG databases

Nicolas Jay, Amedeo Napoli, François Kohler

Cancer Patient Flows Discovery in DRG Databases

Nicolas JAY[a,b1], Amedeo NAPOLI[b] and François KOHLER[a]
[a] *Laboratoire SPI-EAO, Faculty of Medecine., Nancy, France*
[b] *Orpailleur team, Lorraine Laboratory of IT Research and its Applications, Nancy, France*

Abstract. In France, cancer care is evolving to the design of regional networks, so as to coordinate expertise, services and resources allocation. Existing information systems along with data-mining tools can provide better knowledge on the distribution of patient flows. We used one year data of the French Diagnosis Related Groups (DRGs) based system to perform our analysis. Formal Concept Analysis has been used to build Iceberg Lattices of cancer patient flows in the French region of Lorraine. This unsupervised conceptual clustering method allowed us to describe patients flows with an easily understandable visual representation.
Keywords: Patient Flows, Data-Mining, Cancer, Information Systems

1. Introduction

The Cancer Plan, launched by the French government in 2003, is leading to deep changes in healthcare organization. Among measures advocated by this plan: "All establishments providing cancer care shall coordinate their expertise and services, within a set of regional Cancer Poles. Decisions regarding the deployment of major new equipment to be used on a regionwide basis shall systematically be achived through coordination among the main centers concerned... Local networks shall be developed to meet the needs for local coordination."[1].

Such goals cannot be reached without a highly cooperative information system allowing health professionals to share medical information and administrators to manage healthcare organization. But such a system is still in a development stage. The "Programme de Médicalisation des Systèmes d'Information (PMSI)" is the so-called French DRGs based information system. It could be very useful to a better understanding of the actual situation, especially to answer questions about patient

[1] Nicolas Jay. Faculté de Médecine - 9 Avenue de la Forêt de Haye -BP 184 - 54 505 Vandoeuvre Cedex – France. Email : nicolas.jay@medecine.uhp-nancy.fr

flows within the healthcare system.

However it collects large amounts of data and classical statistics methods can be inappropriate to describe their complexity. In that context, Knowledge Discovery tools appear to be a good way of dealing with patient flows.

We propose in this paper a data-mining approach to explore cancer patient flows relying on Formal Concept Analysis, an unsupervised conceptual clustering method providing interesting visualization features[2]. Concept lattices are also known as Galois lattices. Our objective is to describe cancer patient flows in the French region of Lorraine with an easily understandable visual metaphor.

2. Material and Methods

2.1.Materials

Data were extracted from the year 2003 PMSI database of the Lorraine region. We identified hospital stays related to cancer care through the use of an algorithm based on selected codes from the International Classification of Diseases (10th Revision). Since 2001, a cryptographic key can be used in the PMSI to anonymously link successive records of a same patient, whatever the location of the stay. We used this key to rebuild each cancer patient path. Data were then analyzed by subgroup according to fourteen cancer locations (i.e.: breast, lung, nervous system...).

2.2.Methods

Formal Concept Analysis (FCA) is a theory of data analysis identifying conceptual structures within data sets, introduced by Wille [3]. A strong feature of FCA is its capability of producing graphical visualizations of the inherent structures within data. This model 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_{PH} representing I, the relation between a set of four patients P={P1,P2,P3,P4}and a set of four hospitals H={H1,H2,H3,H4}. A cross indicates that a given patient had a stay in the corresponding hospital. A formal concept of K_{PH} is a pair (A,B) with $A \subseteq P$ and $B \subseteq H$ such that (A,B) is maximal with the property $A \times B \subseteq I$. A and B satisfy the following properties:

$$B=\{h \in H / (p,h) \in I \text{ for all } p \in A\} \quad (1)$$

$$A=\{p\in P/\ (p,h)\in I \text{ for all } h\in B\} \quad (2)$$

A is called the intent and B is called the extent of the formal concept. For example, C=({P2,P4},{H1,H2}) is a formal concept of K_{PH}. No patient can be added to the extent of C without modifying its intent. Moreover, no patient can be removed from the extent because in that case, C would not be maximal.

A subconcept - superconcept relation can be formalized as:

$$(A1,B1)\leq(A2,B2) \Leftrightarrow A1\subseteq A2 \Leftrightarrow B2\subseteq B1 \quad (3)$$

Equivalence in (3) comes from (1) and (2). The set of all formal concepts of a formal context K=(G,M,I) together with the order relation $\leq$ is a complete lattice and can be represented in a line diagram as shown in the left part of figure 1 for the K_{PH} formal context.

Table 1: A formal context K_{PH} *representing patients and their hospital stays*

	H1	***H2***	***H3***	***H4***
P1	X		X	X
P2	X	X		
P3		X		
P4	X	X		

Such line diagrams can be very useful in the field of knowledge discovery to understand conceptual relationships within data. A drawback lies in the number of formal concepts that increases significatively (at worst exponentially) with the size of the formal context. Stumme and al. [4] have introduced the notion of iceberg concept lattices. This approach simplifies the line diagram by keeping only the most frequent concepts. Let C=(A,B) a concept of the formal context K=(G,M,I). The support count of C is defined as *supp*(c)=|A|/|G|. Given a minimum support threshold *minsupp*$\in$ [0;1], C is frequent if: *supp*(c)$\geq$*minsupp*. Iceberg lattices can be viewed as filters representing the top-most part of concept lattices. Figure 1 shows the iceberg lattice K_{PH} associated with *minsupp*=0.5

.

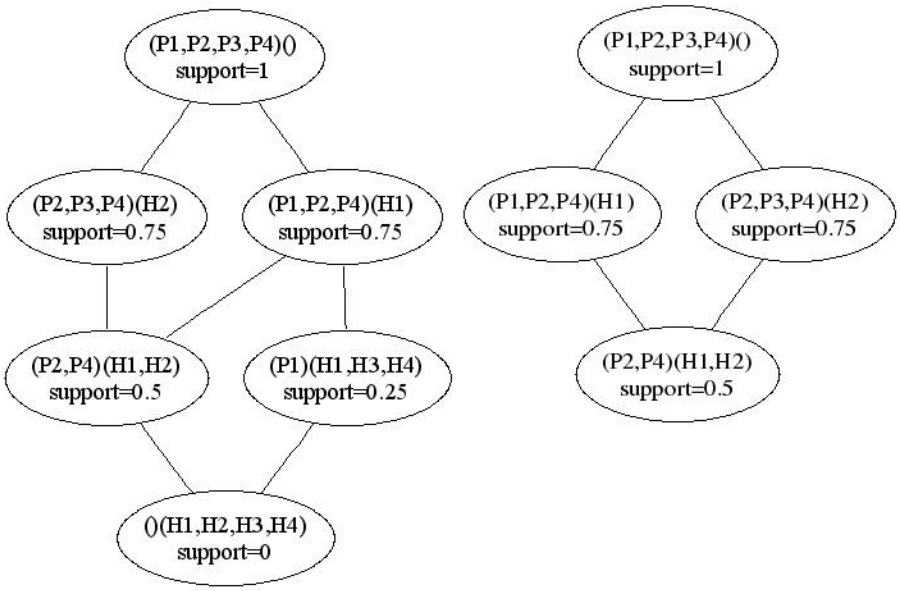
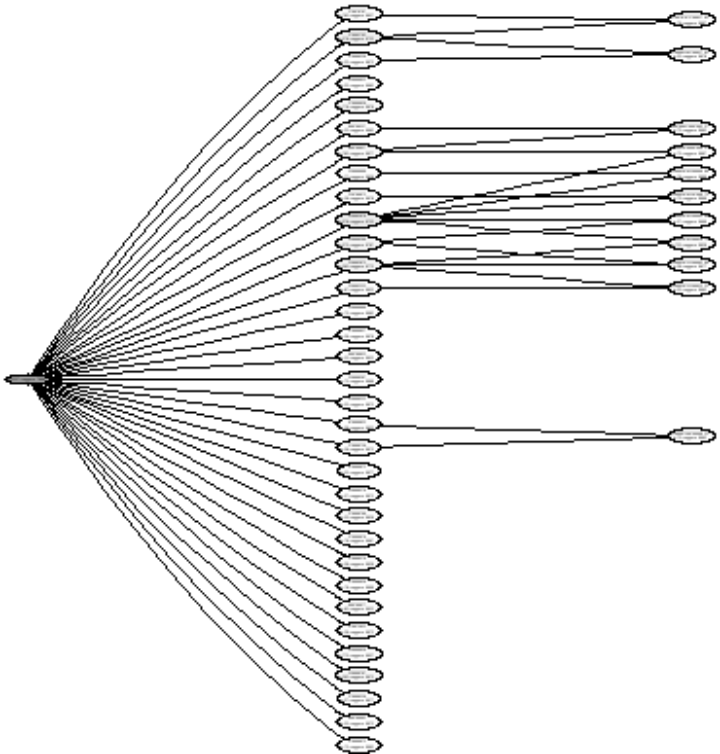


Figure 1: The lattice of the K_{PH} formal context and its related iceberg with minsupp=0.5

In order to discover patterns of hospital care among cancer patients in Lorraine, we built the formal context of the relation between those patients and the hospitals where they had a stay. A specific formal context was created for each cancer location. We chose the Titanic algorithm [5] to extract iceberg lattices from the formal context. Each time, several *minsupp* were tried to find the best suited representation.

3. Results

We propose in this section an example of results that we have obtained with our system. 2257 patients were selected in the original database with a lung cancer. Figure



2 shows the related iceberg lattice with a *minsupp* set at 0.005(around eleven patients). This graph can be read from left to right. The top node is a concept with an empty intent (no hospital) and whose extent contains all the patients. Second layer nodes are concepts with an intent made of one hospital. Third layer nodes illustrate cooperations between hospitals of the second layer nodes. This figure clearly shows four kinds of flows.and may be interpreted as follows. Firstly, a majority of concepts are not connected with any other. Most of the time, they represent hospitals receiving a few patients, but in some cases hospitals being geographically isolated from other major facilities. Secondly, there are three groups of connected concepts. The upper group in the figure concerns hospitals located near Metz, one of the two most populated towns of Lorraine. The middle one shows flows of patients going through two important hospitals located in Nancy, the other big town of Lorraine: The teaching hospital of Nancy (CHU) and the “Centre Alexis Vautrin”, an anti-cancer center (CLCC).

Figure 2: Iceberg Lattice showing flows of patient treated for lung cancer in Lorraine.

In the lower part of the figure, there is a group of three connected nodes. It concerns a hospital located at the Lorraine border, whose patients are also treated in the university teaching hospital of Alsace, the nearby region.

Figure 3 is a zoom of the upper part of the lattice. Each node is greyed according to its support. Labels show hospital identifiers, patient number in the concept's extent, and concept support.

The Nancy CHU (id=540002078) receives the largest number of patients (n=510) ahead of the Metz regional hospital (id=570005165) (n=370), and the Nancy CLCC (id=540003019) (n=304). Cooperation between hospitals can be considered at different levels. Of the 182 patients treated in the Epinal town hospital (id=88078051), only 10 percents (n=19) are shared with the Nancy CHU. On the contrary, half of the patients who have a stay in the Nancy CLCC (144 of 304) have been also treated in the Nancy CHU. These two institutions are the ones sharing the largest number of patients.

Since this iceberg lattice has been built with a 0.005 minimum support, all concepts involving less than 13 patients are pruned from the lattice. This means that even if they exist, smaller flows are not displayed in the lattice.

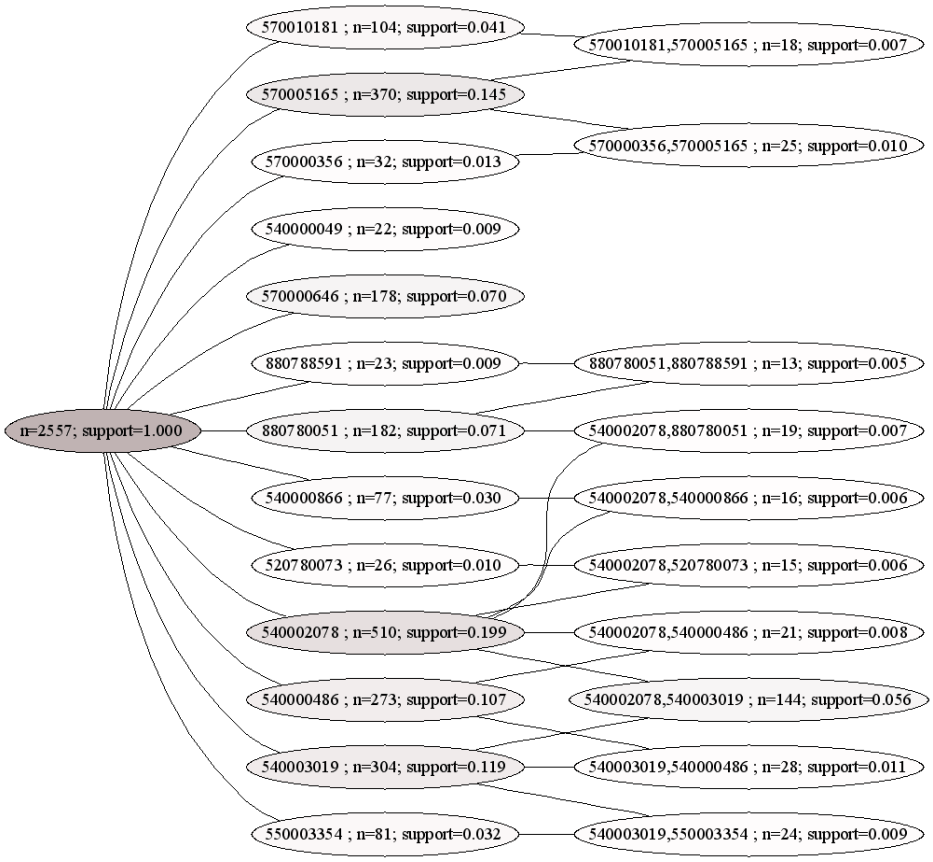


Figure 3: Zoom of the figure 2 lattice. Node labels show hospitals ID of the concept intent, number of patients of the extent, support of the concept.

4. Discussion

Our approach provides several useful features to discover and graphically represent patient flows extracted from a large database. First of all, the classification method is unsupervised and there is no use to consider the exponential number of all possible path combinations through hospitals. The analyst (an expert using the system) handles a threshold to keep the most significant information. Iceberg lattices are a hierarchical clustering method. Thus, they are an interesting alternative to geographical information systems for at least two reasons: they can symbolize flows through more than two sites, and they can highlight flows that do not necessarily rely on a geographical logic. Furthermore, iceberg lattices can be used to compute association rules.

Comparing to a close work [6] our method exploits the same type of data and

relies also on the frequent itemset extraction method. While their concern is more about intra-hospital patient path, we analyzed data for a whole region. We present here a graphical approach providing end-users a global view of patient flows, whose interpretation requires only a minimum of training. We do not here take into account the sequential aspect of the problem, but we believe it is a promising way of research. Iceberg lattices could be used as a pre-processing step to refine the extraction of sequential patterns from patient paths.

Our work can be improved by exploring the nature of the discovered flows: especially when geographical proximity does not explain it, other variables should be added to the formal context. But we will then have to cope with more complex lattices. Another research direction should be to associate iceberg lattices with other classification methods, for example to deal with numeric data.

5. Conclusion

As healthcare networks are in the progress, new tools have to be assessed to deal with the growing amount and complexity of medico-economic data. Data-mining methods can be successfully used to extract new knowledge units from large medical databases. In this work, we are able to describe cancer patient flows at a regional level. This can help healthcare managers to make choices for resource allocation as well as clinicians for cooperative work.

References

[1] Mission Interministérielle pour la Lutte contre le Cancer . http://www.plancancer.fr/ (last accessed Jan. 29, 2006).

[2] Bernhard Ganter, Rudolf Wille, Formal concept analysis, Springer-Verlag, Berlin (DE), 1999.

[3] Wille R, Rival, I. editors. Restructuring Lattice Theory: an approach Based on Hierarchies of concepts Ordered Sets,.Reidel, 1982.

[4] G. Stumme: Efficient Data Mining Based on Formal Concept Analysis. In: A. Hameurlain, R. Cicchetti, R. Traunmüller (Eds.): *Database and Expert Systems Applications*. Proc. DEXA 2002. LNCS 2453, Springer, Heidelberg 2002, 534-546.

[5] G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, L. Lakhal: Computing Iceberg Concept Lattices with Titanic. *J. on Knowledge and Data Engineering (*KDE) 42(2), 2002, 189-222.

[6] Dart T, Cui Y, Chatellier G, Degoulet P. Analysis of hospitalised patient flows using data-mining. Stud Health Technol Inform. 2003;95:263-8.