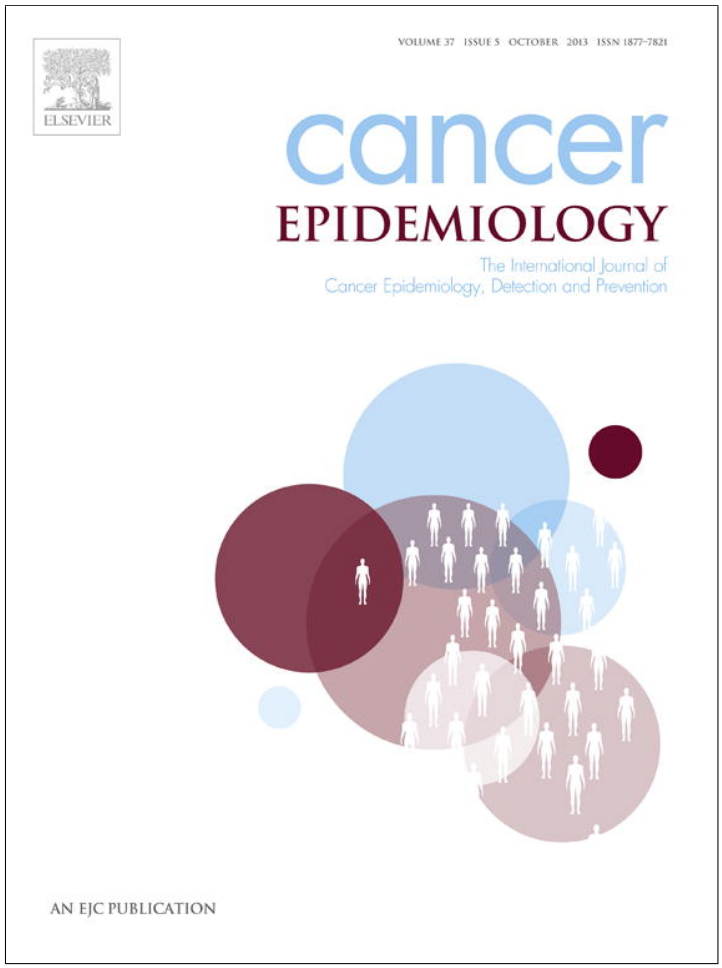

Cancer Epidemiology 37 (2013) 688–696

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 healthcare 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

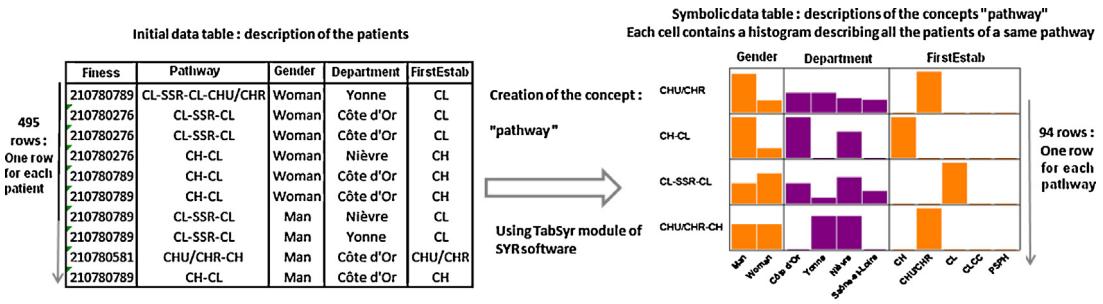

Fig. 1. From the patient's data table to the pathways data table using symbolic data analysis methodology and the SYR software.

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.

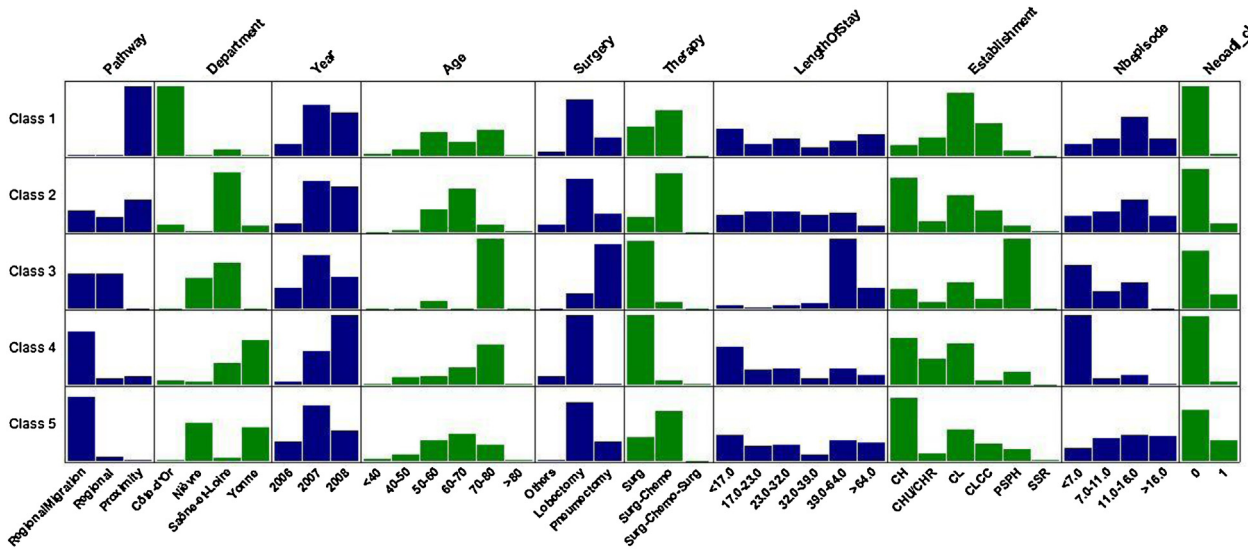

Fig. 2. Description of the 5 classes (profiles) using pathway description variables.

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

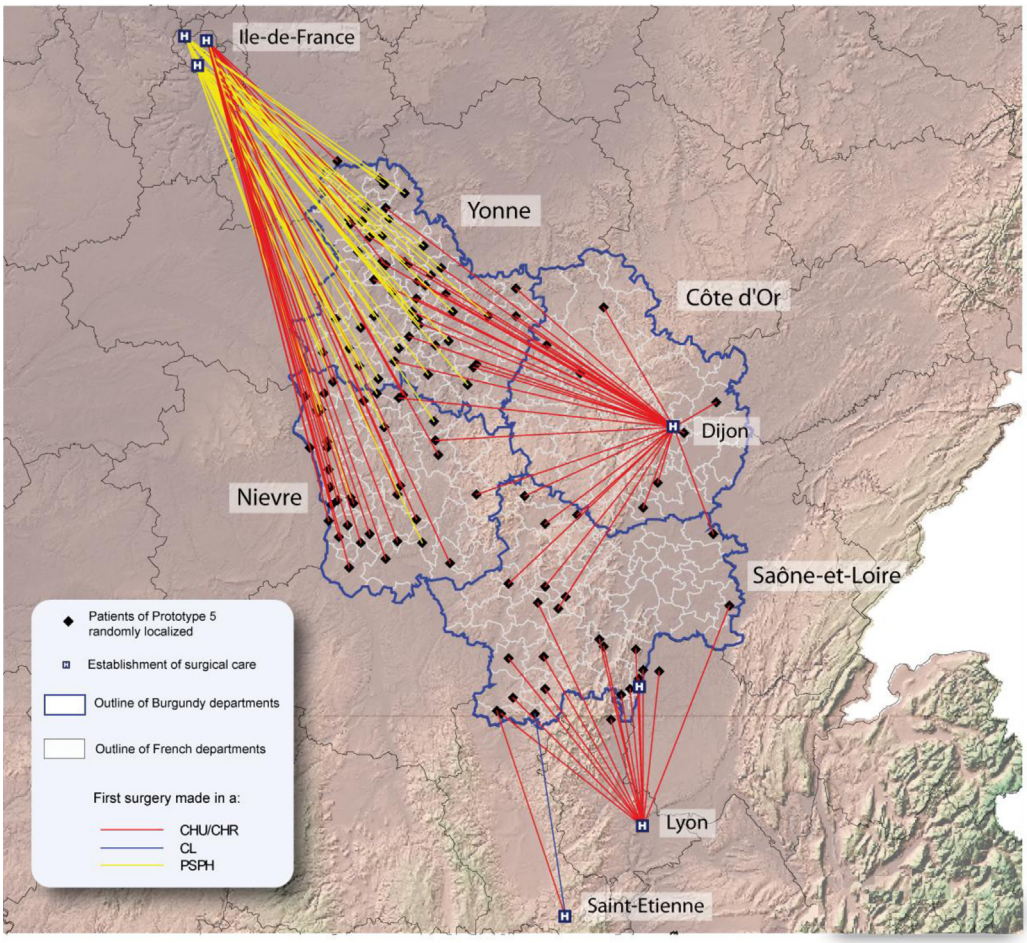

Fig. 3. Cartographic representation of class 5.

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.