Identifying homogeneous healthcare use profiles and treatment sequences by combining sequence pattern mining with care trajectory clustering in kidney cancer patients on oral anticancer drugs: A case study

Cyril Baudrier, Yohann Tran, Nicolas Delanoy, Sandrine Katsahian, Brigitte Sabatier, Germain Perrin

Original Research Article

Health Informatics Journal

Health Informatics Journal
1–18

DOI: 10.1177/14604582221101526
journals.sagepub.com/home/jhi

Identifying homogeneous healthcare use profiles and treatment sequences by combining sequence pattern mining with care trajectory clustering in kidney cancer patients on oral anticancer drugs: A case study

Cyril Baudrier
Pharmacy Department, Hospital European Georges Pompidou, Paris, FR

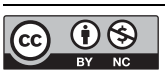
Yohann Tran
Clinical Research Department, Hospital European Georges Pompidou, Paris, FR

Nicolas Delanoy
Oncology Department, Hospital European Georges Pompidou, Paris, FR

Sandrine Katsahian
Clinical Research Department, Hospital European Georges Pompidou, Paris, FR;
Cordeliers Research Centre, INSERM, Paris, FR;
Inria, HeKA, Paris, FR

Brigitte Sabatier and Germain Perrin
Pharmacy Department, Hospital European Georges Pompidou, Paris, FR;
Cordeliers Research Centre, INSERM, Paris, FR;
Inria, HeKA, Paris, FR

Corresponding author:
Cyril Baudrier, Pharmacy Department, Georges Pompidou European Hospital, Assistance Publique- Hôpitaux de Paris, 20 rue Leblanc, Paris 75015, France.
Email: cyril.baudrier@aphp.fr



Abstract

Objective: We evaluated the ability of a coupled pattern-mining and clustering method to identify homogeneous groups of subjects in terms of healthcare resource use, prognosis and treatment sequences, in renal cancer patients beginning oral anticancer treatment.

Methods: Data were retrieved from the permanent sample of the French medico-administrative database. We applied the CP-SPAM algorithm for pattern mining to healthcare use sequences, followed by hierarchical clustering on principal components (HCPC).

Results and conclusion: We identified 127 individuals with renal cancer with a first reimbursement of an oral anticancer drug between 2010 and 2017. Clustering identified three groups of subjects, and discrimination between these groups was good. These clusters differed significantly in terms of mortality at six and 12 months, and medical follow-up profile (predominantly outpatient or inpatient care, biological monitoring, reimbursement of supportive care drugs). This case study highlights the potential utility of applying sequence-mining algorithms to a large range of healthcare reimbursement data, to identify groups of subjects homogeneous in terms of their care pathways and medical behaviors.

Keywords

Antineoplastic agents, cluster analysis, data mining, kidney neoplasms, patient care management, prognosis, sequential pattern mining

Introduction

Since the approval of sorafenib by the FDA in 2005, oral multikinase inhibitors have become the cornerstone treatment for metastatic renal cell carcinoma (mRCC). The 2021 Updated European Association of Urology Guidelines on Renal Cell Carcinoma recommend combinations of immune checkpoint inhibitors plus oral tyrosine kinase inhibitors (TKI) for the first-line treatment of mRCC. This approach has yielded substantial gains in terms of progression-free and overall survival (OS), relative to oral TKIs alone. Oral TKI monotherapy is still considered for patients unable to take or to tolerate checkpoint inhibitors, and for those displaying no response to immunotherapy.[1] Oral TKI treatment sequences in mRCC patients have been little studied[2] and may constitute prognostic hallmarks of disease progression.[3,4]

The frequency and severity of adverse effects associated with TKI treatment remain high in comparison with intravenous chemotherapies, creating a major challenge in the ambulatory management of mRCC patients, whose healthcare pathways involve both hospital and community healthcare professionals.[5,6] Oral TKI have been associated with multiple degradations or failures in care pathways, particularly at the drug dispensing and administration stages.[5]

In France, patients' healthcare pathways is coordinated by general practitioners, but a recent report revealed that patients are generally dissatisfied with their care pathways.[7] In this context, the 2014-2019 Cancer Plan supported the development of organizational initiatives to ensure good management of patients on oral chemotherapy, by improving the prevention and early management of TKI toxicities through a better cooperation between community and hospital healthcare professionals.[6] An understanding of healthcare trajectories is essential for healthcare planning and optimal allocation of resources, but few data are available concerning healthcare resource use by mRCC patients on TKI, making the assessment of the quality of clinical management difficult.

This study aimed to determine whether a combination of pattern identification in care trajectory sequences and sequence clustering methods could be applied to data from healthcare reimbursement databases to identify mRCC patients with homogeneous care trajectories, particularly in terms of oral TKI treatment sequences and prognosis.

Methods

The results of the study are reported according to the STROBE guidelines.[8]

Design

We performed a retrospective descriptive cohort study on healthcare use during the year following the initiation of oral TKI treatment, in patients with mRCC.

Data source

Ambulatory healthcare data were retrieved from a representative French healthcare database (EGB: *Échantillon Généraliste des Bénéficiaires*) covering 1/97th of the nationwide healthcare insurance database (SNIIRAM: *Système National d'Information Inter-Régime de l'Assurance Maladies,* collecting data from 66 million people, i.e. more than 97% of the French population). The EGB database is representative of the French population (random selection of beneficiaries) regarding age (five-year increments), sex and healthcare expenditures per beneficiary .[9] Each patient is identified with a unique anonymous number in the database. Ambulatory data from the EGB are merged with hospital diagnosis through the *Programme de Médicalisation des Systèmes d'Information* (PMSI).

Identification of renal cancer patients

All renal cancer patients were identified in the EGB database between 1 January 2010 and 31 December 2017 with a major long-term illness (*Affection Longue Durée*, ALD) associated with an ICD-10 code for renal cancer (C64, C65, C66, C88 or D091), with a starting date in the same year or the preceding year (year *n* or *n*-1), and/or patients with a hospital discharge diagnosis of renal cancer during year *n* or *n*-1 (ICD-10 codes same ICD-10 codes as a major or related diagnosis).[10,11] Comorbidity was assessed with the Charlson comorbidity index, based on hospital discharge diagnoses before the index date for TKI initiation.[12]

Extraction of healthcare use data

We identified reimbursements for oral anticancer drug approved for the treatment of mRCC (i.e. sunitinib, axitinib, cabozantinib, everolimus, or pazopanib) based on *Code Identifiant la Présentation* 13 (CIP-13) codes. The one-year follow-up period began on the data of the first reimbursement for an oral TKI. For healthcare service use, we extracted, for each subject, the number of outpatient visits to a general practitioner and specialists, the number of visits to hospital physicians and the number of admissions to the day hospital and emergency department. We calculated the Bice-Boxerman Continuity of Care Index (COCI) to assess the level of dispersion of appointments between different professionals (considering general practitioners, community specialists and hospital specialists). This index ranges from a minimum of 0 to a maximum of 1 (all appointments with the same professional during the follow-up) and is calculated as follows[13]:

$$COCI = \frac{\left(\sum_{i=1}^{n} n_i^2\right) - n}{n(n-1)}$$

where n = total number of consultations, n_i = number of consultations with the i^{th} healthcare provider (from 1 to 3).

We also extracted reimbursements of drugs prescribed for the treatment of adverse effects associated with TKI or associated with disease progression: antinausea (ATC class: A04), antidiarrhea (A07), antihypertensive (C02), and opiates (N02A). Reimbursement data for the biological monitoring of TKI therapy, as recommended in national guidelines, were also extracted based on NGAB codes, including renal (NGAB codes: 592, 2004) and hepatic (codes: 514, 516, 517, 519, 1601) function evaluation and total blood counts (TBC, codes: 1104). For each subject, we calculated the medication possession ratio (MPR), corresponding to the total number of days of TKI treatment collected at the community pharmacy, divided by the number of days of follow-up (number of days between TKI initiation and end of follow-up or death).[14]

Pattern mining in care trajectories

Frequent care sequences were identified from each individual chronologically ordered care trajectory with a sequential pattern-mining algorithm. We considered visits to general practitioners and community specialists, and visits to hospital physicians, emergency department and admissions to day hospital in the construction of care trajectories. Sequential pattern mining considers the order of each element in the sequence. A frequent sequence is defined as a string of characters appearing recurrently in a dataset, at a frequency higher than a fixed minimum support threshold. We used the contextual sequential pattern mining (CM-SPAM) algorithm to identify frequent sequences.[15] This algorithm was executed with SPMF (v.2.42), with a support threshold of 30% and a maximum gap of 1, to study consecutive care events.

Clustering models

Clusters were identified by hierarchical principal component classification, on the basis of the frequent sequences identified by CM-SPAM. We used a mixed dataset composed of continuous and categorical (including all frequent sequences) variables, making utilization of factor analysis of mixed data (FAMD) approach appropriate. FAMD is a clustering method used to summarize a dataset through a main axis, corresponding to a linear combination of variables.[16] Vital status at 6 and 12 months was added as additional variables. The choice of the optimal number of dimensions explaining our dataset was based on the elbow rule: interpretable results with the maximum observed inertia and the minimum factor. We then performed hierarchical clustering on principal components (HCPC) based on the FAMD results.[17] For each subject, we calculated a score quantifying the similarity between the subject's follow-up sequence and each of the frequent sequences identified. A score of 1 was attributed if the follow-up sequence was identical to the frequent sequence, and a score of 0 otherwise.

Sensitivity analysis

We first performed a cross-validation using 4 random subsamples equivalent to approximatively 75% of the complete database. We evaluated the variation in the number of sequences identified and

the number of subjects in each cluster. Second, we modified the support from 30% (main analysis) to 70% to test the variation in the number of sequences identified and the proportion of subjects reclassified. Finally, we used K-means clustering algorithm with a prespecified number of 3 clusters as an alternative to HCPC method, to test the variation in the subject cluster assignment. Results of the sensitivity analysis are given in Supplemental material 1.

Statistical analysis

Qualitative variables are expressed as numbers and associated percentages, and quantitative variables are expressed as the mean ± SD or median and interquartile range. Groups were compared in parametric Student's *t* test or non-parametric Mann-Whitney-Wilcoxon/Kruskal Wallis tests for quantitative variables, depending on variable distribution, and with Chi-squared tests/Fisher's test for qualitative variables. A $p < 0.05$ were considered statistically significant. Statistical analysis was performed with R-Studio software (version 1.4.1106).

Ethics and data protection

Access to the database is legally authorized without the need for permission from the national data protection agency (CNIL).

Results

Characteristics of the population

We identified 1442 individuals with renal cancer in the EGB database from 2010 to 2018. At least one oral TKI reimbursement was recorded in the EGB database for 127 of these individuals (8.8%). A flowchart of the study is provided in Figure 1. Mean age was 65.2 ± 10.7 years (range: 32 to 86 years), and most subjects were men (70.1%), which corresponded to the epidemiology of renal cancer in France.[18]

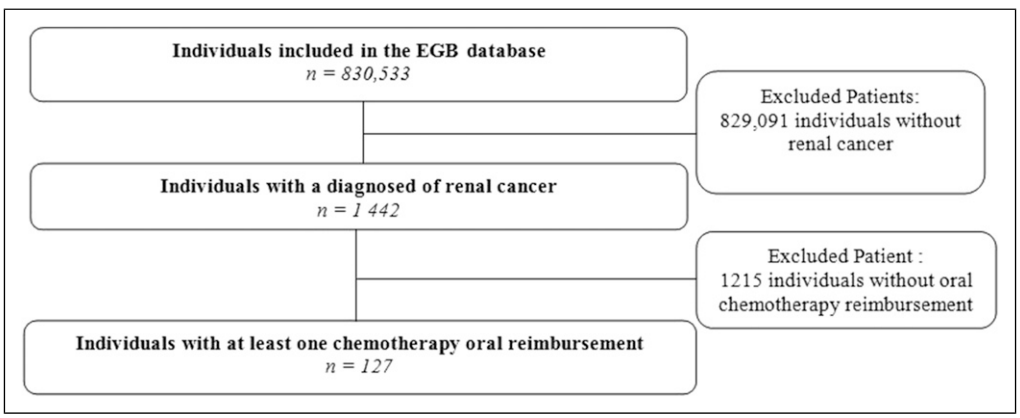


Figure 1. Flowchart of the study.

Mean follow-up duration was 299 ± 112.9 days during the first year after the oral TKI initiation. Twenty-five (19.7%) subjects died by six months, and 40 (31.5%) died by one year after treatment initiation. The mean number of long-term diseases per patient was 1.4 ± 0.8.

Pattern mining

We identified 120 frequent sequences using the CM-SPAM algorithm, with 30% support and a maximum gap of 1. The list of sequences and associated supports is provided in the Supplemental material 2.

Cluster identification

We identified homogeneous healthcare use groups with a clustering method based on the frequent sequences identified by the algorithm. The optimal number of dimensions explaining our data set well with interpretable results, maximum observed inertia and the minimum factor was 3 (Scree Plot is given in Figure 2.). The three axes obtained accounted for 42.9% of the variability observed in our dataset.

We then applied HCPC to the FAMD results. Axes 1 and 2 discriminated three homogeneous clusters of subjects well, as shown in Figure 3 and Figure 4 (distribution of the 120 frequent sequences is given in Supplemental material 3). Clusters were relatively balanced, with 42 subjects in cluster 1, 58 in cluster 2 and 27 in cluster 3.

Patients and healthcare use profiles in each cluster

The results obtained for healthcare use for the total sample and for each cluster are presented in Table 1.

We found that cluster 1 ($n = 42$) was associated with a mean follow-up of 310.4 ± 107.8 days, a 6-`month mortality of 16.7% and a significantly higher rate of access to outpatient care during follow-up (median of 2.7 visits/100 days to a general practitioner, and 3.4 visits/100 days to a community specialist) than for clusters 2+3, and a lower rate of access to hospital practitioners than the total population. In cluster 1, 18 (42.9%) subjects visited the emergency department, and 18 subjects (42.9%) were admitted to the day hospital. The continuity of care index (COCI) was 0.47. The proportion of subjects initiating an antihypertensive treatment after TKI initiation was smaller for cluster 1 (4.8%) than for clusters 2+3. For the first TKI prescribed during the one-year follow-up period, 13 (31%) subjects experienced a dose reduction, after a median of 49 days [39-93 days]. The median MPR was 78% for the first year of TKI treatment. This value is lower than that for clusters 2+3, but not significantly so.

Subjects in cluster 2 tended to be younger (63.7 ± 10.4, $p = 0.1$ versus cluster 1+3), and comorbidity rate was lower in cluster 2 than in clusters 1+3 (Charlson index 7.1 ± 3.6). Follow-up was longer in cluster 2, with a mean value of 328.1 ± 89.2 days. Cluster 2 also had the lowest 6-month (8.6%) and 1-year (19.8%) mortality rates. We found that 43.1% of the subjects attended the emergency department at least once during the follow-up, and 29.3% was admitted to the day hospital. This cluster was associated with a higher rate of referral to hospital specialists and a lower rate of outpatient visits to community practitioners. The COCI was lower (0.37) than that of cluster 1+3, reflecting visits to a greater diversity of practitioners during follow-up. The frequency of dose reduction for the first TKI prescribed was similar to that for cluster 1, but the time to first

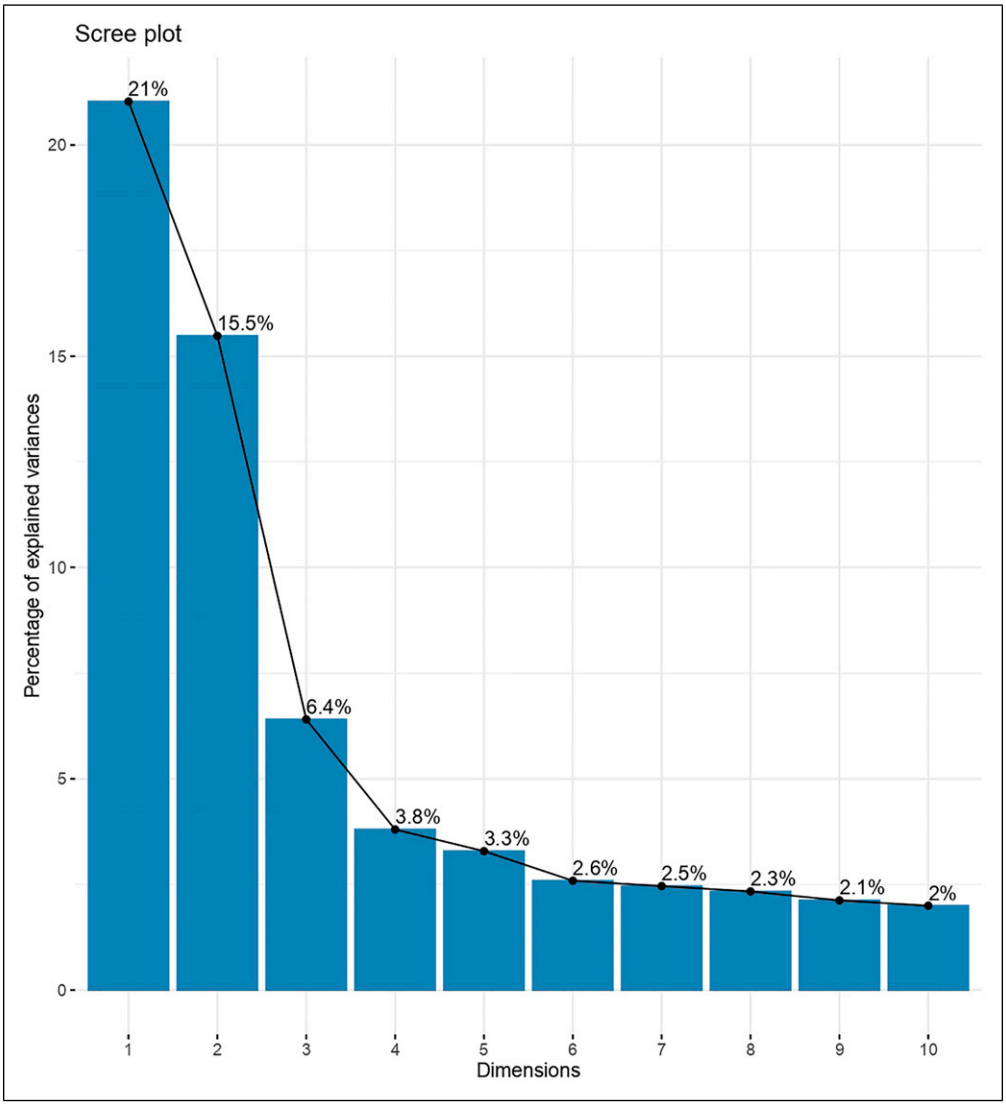


Figure 2. Scree plot.

dose reduction was longer (although not significantly so), at 89.5 days. In this cluster, the rate of ambulatory blood and renal surveillance was significantly lower than that in clusters 1+3.

Cluster 3 had the shortest follow-up, of 221.3 ± 132.9 days, and the highest 6-month and 1-year mortality rates (48.2% and 66.7%, respectively), with a low utilization of community practitioners and a higher rate of referral to hospital practitioners. The COCI index was significantly higher than that of clusters 1+2 (0.62), revealing visits to a limited diversity of physicians during the follow-up. Finally, this cluster was associated with a higher rate of ambulatory biological monitoring, which was significant for blood monitoring.

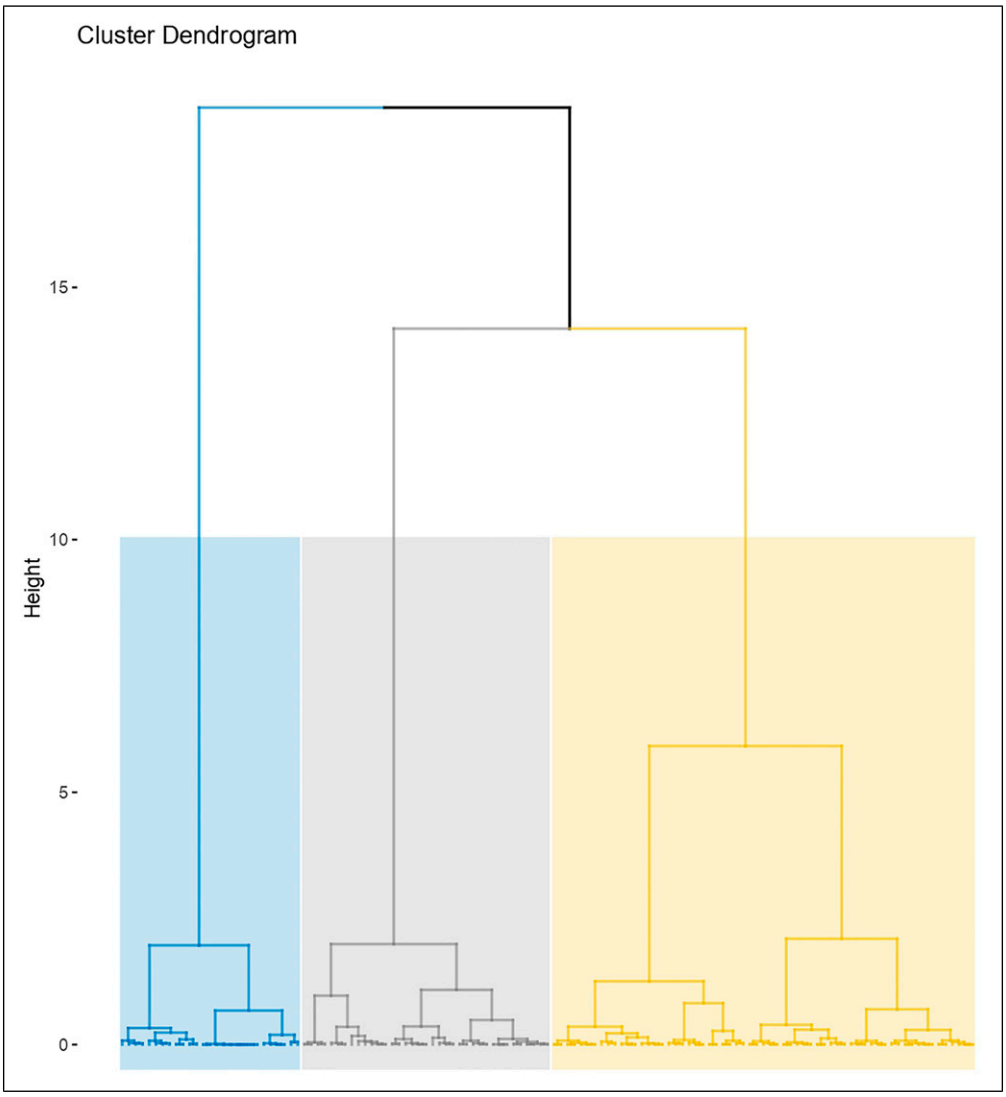


Figure 3. Cluster dendrogram showing three distinct homogeneous clusters of subjects.

The proportion of subjects with reimbursements for supportive care treatments (i.e. anti-nausea, antidiarrhea and opiates) did not differ significantly between clusters and ranged from 28.4% to 38.6%.

TKI sequences

In all clusters, the main first-line therapy was sunitinib. The rate of maintenance of this therapy (without considering possible dose reductions) at six months and one year was high in cluster 2, intermediate in cluster 1, and low in cluster 3 (for the one-year maintenance rates—47.5%, 33.3%

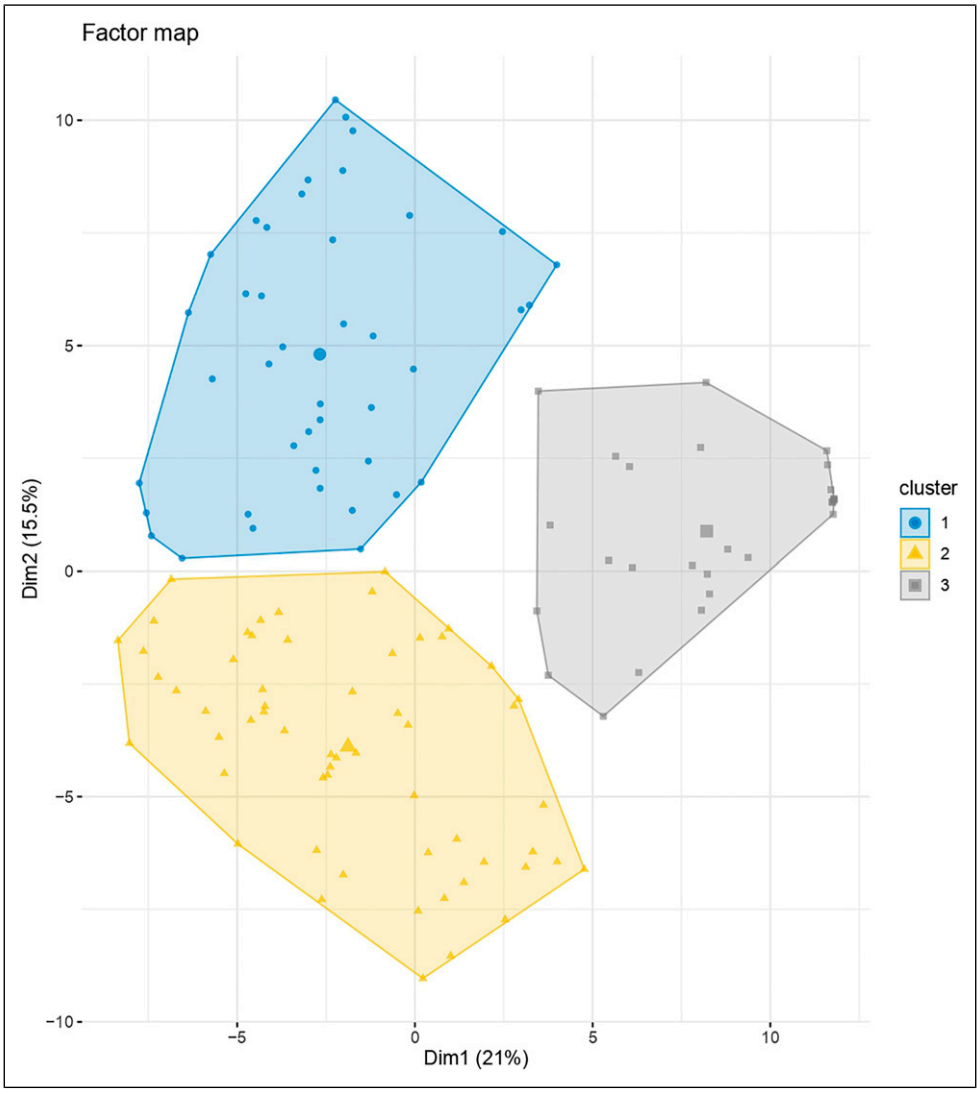


Figure 4. Factor map showing the distribution of the 127 subjects according to axes 1 and 2.

and 4.5% respectively—see Figure 5). Pazopanib was the second most frequent first-line treatment in clusters 1 and 2, with a 1-year maintenance rate of 57.1% and 61.5%, respectively; but this treatment was not found in cluster 3. Axitinib and sorafenib were also identified as minor first-line treatment choices, as was everolimus in clusters 1 and 3. At the end of follow-up, cabozantinib was also found (cluster 2). In cluster 3, only 11.1% of subjects were still receiving reimbursements for an oral anticancer drug, versus 50% of those in cluster 1, and 65.5% of those in cluster 2. At six months, the rate of continuation of the first treatment was 42.8% for cluster 1, 58.6% for cluster 2 and 11.1% for cluster 3. At one year, this rate was 33.3% for cluster 1, 41.4% for cluster 2 and 7.4% for cluster 3.

Table 1. Healthcare use profiles in clusters 1, 2 and 3.

	Total (*n* = 127)	Cluster 1 (*n* = 42)	Cluster 2 (n = 58)	Cluster 3 (n =27)	*p*-value cluster 1 vs. clusters 2-3	*p*-value cluster 2 vs. clusters 1-3	*p*-value cluster 3 vs. clusters 1-2
Patients characteristics							
Age, mean ± SD	65.2 ± 10.7	67.2 ± 9.7	63.7± 10.4	65.5 ± 12.4	0.2	0.1	0.8
Female, *n* (%)	38 (29.9)	10 (23.8)	19 (32.8)	9 (33.3)	0.3	0.5	0.7
Number of long-term diseases, mean ± SD	1.4 ± 0.8	1.4 ± 0.8	1.5 ± 0.9	1.2 ± 0.8	0.9	0.3	0.3
Charlson comorbidity index, mean ± SD	7.6 ± 3.2	8 ± 2.7	7.1 ± 3.6	7.9 ± 3.6	0.5	0.3	0.6
History of diabetes, *n* (%)	42 (33.1)	9 (21.4)	20 (34.5)	13 (48.2)	0.05	0.8	0.06
History of renal disease, *n* (%)	13 (10.2)	5 (11.9)	7 (12.1)	1 (3.7)	0.9	0.7	0.4
History of malnutrition, *n* (%)	34 (26.8)	12 (28.6)	15 (25.9)	7 (25.9)	0.7	0.8	0.9
Days of follow-up, mean ± SD	299.5 ± 112.9	310.4 ± 107.8	328.1 ± 89.2	221.3 ± 132.9	0.4	0.005	<0.001
Death at M_6, *n* (%)	25 (19.7)	7 (16.7)	5 (8.6)	13 (48.2)	0.5	0.004	<0.001
Death at M_{12}, *n* (%)	40 (31.5)	11 (26.2)	11 (19.8)	18 (66.7)	0.4	0.005	<0.001
Access to healthcare professionals							
General practitioner visits/100 days, median [IQR]	1.9 [0.8-3.6]	2.7 [1.8-3.8]	1.9 [0.8-4.0]	0.3 [0.0-2.3]	0.003	0.8	<0.001
Specialist visits/100 days, median [IQR]	2.2 [0.5-3.6]	3.4 [2.5-3.8]	1.8 [0.6-2.8]	0.0 [0.0-1.0]	<0.001	0.6	<0.001
≥1 ED visit, *n* (%)	51 (40.2)	18 (42.9)	25 (43.1)	8 (29.6)	0.7	0.5	0.2
≥1 admission to the day hospital, *n* (%)	42 (33.1)	18 (42.9)	17 (29.3)	7 (25.9)	0.1	0.4	0.4
Hospital specialist visits/100 days	1.9 [0.3-3.3]	0.1 [0.0-1.0]	2.7 [2.0-4.2]	2.2 [0.3-3.7]	<0.001	<0.001	0.5
Bice-Boxerman continuity of care index (COCI), median [IQR]	0.45 [0.35-0.50]	0.47 [0.42-0.50]	0.37 [0.33-0.43]	0.62 [0.49-0.95]	0.2	<0.001	<0.001

(continued)

Table 1. (continued)

	Total (n = 127)	Cluster 1 (n = 42)	Cluster 2 (n = 58)	Cluster 3 (n =27)	p-value cluster 1 vs. clusters 2-3	p-value cluster 2 vs. clusters 1-3	p-value cluster 3 vs. clusters 1-2
Use of medication							
Number of oral antitumor treatments, median [IQR]	1.0 [1.0-2.0]	1.0 [1.0-2.0]	1.0 [1.0-1.75]	1.0 [1.0-1.5]	0.6	0.6	0.9
Dose reduction for the first oral antitumor treatment, n (%)	34 (26.8)	13 (31.0)	18 (31.0)	3 (11.1)	0.5	0.3	0.04
Time to dose reduction for the first antitumor treatment in days, median [IQR]	64 [40.2-120.0]	49 [39-93]	89.5 [53.3-135.0]	34 [23-38]	0.4	0.3	0.6
Medication possession ratio for oral antitumoral treatment, median [IQR]	90.1 [59.2-100.0]	78.0 [47.0-98.2]	92.1 [65.8-100]	94.4 [79.5-100.0]	0.7	0.6	0.3
Anti hypertensive drug initiation during follow-up, n (%)	22 (17.3)	2 (4.8)	16 (27.6)	4 (14.8)	0.02	0.01	0.9
1 box of antinausea medication reimbursed, n (%)	36 (28.4)	13 (31.0)	17 (29.3)	6 (22.2)	0.6	0.8	0.4
1 box of antidiarrhea medication reimbursed, n (%)	49 (38.6)	13 (31.0)	25 (43.1)	11 (40.7)	0.2	0.3	0.8
1 box of an opiate drug reimbursed, n (%)	47 (37.0)	15 (35.7)	24 (41.4)	8 (29.6)	0.8	0.3	0.4
1 box of anti hypertensive medication reimbursed, n (%)	43 (33.9)	27 (64.3)	42 (72.4)	15 (55.6)	0.8	0.2	0.2
Biological monitoring							
Total blood counts/100 days, median [IQR]	4.4 [2.7-6.7]	4.3 [2.7-6.8]	3.6 [2.3-5.2]	7.2 [3.3-12.0]	0.7	0.01	0.008
Renal function assessments/100 days, median [IQR]	1.1 [0.0-2.6]	1.9 [0.6-3.0]	0.5 [0.0-1.9]	1.4 [0.0-3.1]	0.006	0.006	0.8
Hepatic function assessments/100 days; median [IQR]	7.4 [3.6-12.1]	7.4 [4.2-11.5]	7.8 [3.9-11.5]	7.5 [2.3-13.1]	0.9	0.9	0.8

COCI: Continuity of Care Index, ED: emergency department, IQR: interquartile range, SD: standard deviation.

Sensitivity analysis

The cross-validation in 4 different random subsamples showed that the variation in number of frequent sequences identified between subsamples was low (i.e. ≤ 10%), and the proportion of subjects reclassified in another cluster was <3.5%.

When applying a K-means clustering algorithm with a predefined number of three clusters instead of HCPC, we found that all the regions maintained the same allocation of hierarchical clustering (see Supplemental material 1).

Discussion

This is, to our knowledge, the first study of the use of the CM-SPAM algorithm coupled with clustering methods to depict the care trajectories of cancer patients included in a healthcare reimbursement database. Data mining approaches have already been used in French healthcare databases, to study care trajectories in the context of breast cancer with a formal concept analysis,[19] prenatal care consumption with state sequence analysis[20] or acute coronary syndrome with contextual frequent pattern mining.[21] We hypothesized that the order of consecutive care in the patient sequence would be crucial for the identification of hallmarks of quality of care. We therefore used a sequential pattern mining method, with restrictive rules in terms of support (we selected sequences found at least in 30% of subjects) and gaps (we investigated only sequences of consecutive cares). We also created a matrix of similarity index, assigning a coefficient of 1 if a given sequence was found in a patient's care trajectory, and 0 otherwise, to compare patients with frequent sequences with more weight given to individuals following strictly identical pathways in terms of frequent sequences. Since sequences were long and dense, we used the CM-SPAM algorithm to identify frequent sequences. This algorithm works by vertical extraction of sequential patterns, rendering it faster and less expensive than the SPAM method.[15]

Our approach, based on a combination of CM-SPAM with FAMD, identified three groups that were homogeneous in terms of care trajectories and patient prognosis. Groups can be flagged as follow: (i) predominant ambulatory follow-up and biological monitoring, with intermediate mortality rate at 6 and 12 months, (ii) mixed ambulatory and hospital follow-up, with lower rate of ambulatory biological monitoring, with low 6- and 12-month mortality rates and (iii) predominant hospital follow-up, with a poor 6- and 12-month prognosis, and a higher continuity of care. FAMD dimension reduction is an important tool for transforming complex mixed data into lower-dimensional subspaces while preserving important characteristics of the original data. This technique is useful to reduce complexity and support decision making.[22] Discriminant characteristics of each cluster are summarized in Table 2.

We found that continuity of care can be conveniently integrated in analysis of healthcare trajectories through COCI calculation. In our study, we found that cluster 3 was associated with a higher COCI as compared with cluster 1+3. Previous studies showed that COCI was correlated with a lower requirement for ED services among multiple chronic conditions patients, supporting our observations.[23]

Concerning ambulatory biological monitoring, our results indicated a relatively adequate biological surveillance in the population, since French learned societies recommend a three-monthly clinicobiological evaluation for mRCC patients, including TBC and renal function assessments (approximatively 1 biological test per 100 days).[24] We found differences between clusters for the number of TBC/100 days, with cluster 3 being associated with the highest rate, probably in connection with the poor 6-month and 12-month prognosis in this cluster. Integrating

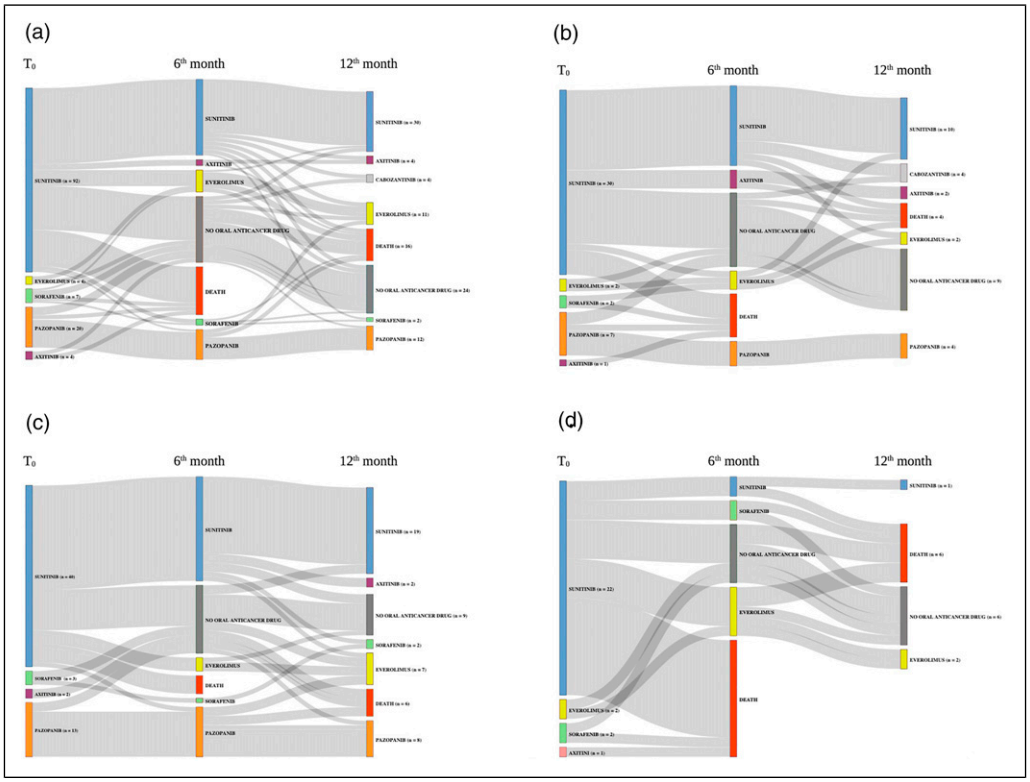


Figure 5. Treatment sequences between T0 (initiation of the first oral anticancer drug), 6th month and 12th month, in the total sample (panel A), in cluster 1 (panel B), in cluster 2 (panel C) and in cluster 3 (panel D).

biological data in care trajectory analysis can provide useful hallmarks of disease evolution. For example, Ben-Assuli et al. found that clustering patients according to serum creatinine levels trajectory can be an indicator of serious illness resulting in multiple ED visits[25]. We found that cluster 3 was associated with both a higher rate of TBC assessment, and a non-significant lower rate of ED visit.

Beside these discriminant characteristics, we also found general patterns of suboptimal management, which were uniformly distributed between clusters. First, emergency department visit was a frequently identified pattern, since we found that 40.2% of subjects attended the emergency department at least once in the year following TKI initiation. Emergency department visits, and preventable emergency department visits in particular, are important hallmarks of poor-quality management of the patient or inadequate access to healthcare.[26]

Second, our results revealed that only one third of subjects received a reimbursement for drugs prescribed to treat adverse effects associated with oral TKI (i.e. antinausea and antidiarrhea), although anticipatory prescriptions are advised for all patients, to ensure that adverse effects related to TKI treatment are rapidly detected and resolved, to prevent unnecessary ED visits and and to improve patient quality of life. This low rate of reimbursement constitutes a hallmark of suboptimal management, which can be targeted for the implementation of health interventions. We also observed that cluster 2 (best prognosis) was associated with a higher rate of initiation of an

Table 2. Discriminant characteristics between cluster 1, 2 and 3.

	Cluster 1 (*n* = 33.0%) Intermediate prognosis Ambulatory follow-up	Cluster 2 (*45.7%*) Best prognosis Mixed ambulatory and hospital follow-up	Cluster 3 (21.3%) Poor prognosis Close hospital follow-up
Patients characteristics			
Days of follow-up	++	**+++**	+
Death at M_6	++	+	**+++**
Death at M_{12}	++	+	**+++**
Access to healthcare professionals			
General practitioner visits/100 days, median [IQR]	**+++**	++	+
Specialist visits/100 days, median [IQR]	**+++**	++	+
Hospital specialist visits/100 days	+	**+++**	**++**
Bice-Boxerman continuity of care index (COCI), median [IQR]	++	+	**+++**
Use of medication			
Antihypertensive drug initiation during follow-up, *n* (%)	+	**+++**	++
Biological monitoring			
Total blood counts/100 days, median [IQR]	++	+	**+++**
Renal function assessments/100 days, median [IQR]	++	+	++

COCI: Continuity of Care Index, IQR: Interquartile Range.

antihypertensive drug, which can indicate a better management of incident hypertension, which is a specific toxicity related to TKI treatment.[27]

Third, we observed that the MPR was high in our study sample (90%), in accordance with literature.[28] Even though this measure is used as an indirect estimation of medication adherence (i.e. whether patient took the medications as prescribed), such indicator, derived from electronic database, should be interpreted cautiously outside the context of daily clinical practice, since MPR decrease can be related to TKI high-grade toxicity leading to temporary drug discontinuation, even in the presence of a patient with good medication adherence. In this context, the use of alternative measures, such as the relative drug intensity (i.e. the amount of drug administered per unit of time divided by the amount recommended) also associated with patient outcomes including survival, should be preferred.[29]

Altogether, these results could offer operational perspectives for policymakers to tailor healthcare interventions aiming to secure the management of mRCC patient taking oral TKI. Such intervention should target the reduction of ED visits, by improving continuity of care with a better coordination between community and hospital healthcare providers. Additionally, efforts should be made to improve patient medication adherence and/or dose intensity, as well as to improve patient counselling about management of toxicities. In this perspective, recent initiatives have been implemented in France, such as the national experimentation Onco'Link 2021-2024, which allows the

coordination between community and hospital healthcare providers through regular multidisciplinary hospital appointments in day hospitalizations.[30]

In terms of treatment sequences, most of the patients in all three clusters were initiating sunitinib. A pattern of switching to axitinib was observed at six months in cluster 1 but not in cluster 2. Rates of first drug continuation at six months and one year were higher for cluster 2 than for cluster 1, consistent with the lower death rate in this cluster. So far, TKI treatment sequences of mRCC patients has been poorly evaluated, even though this information can provide insights about patient prognosis. Indeed, Finek *et al.* found that TKI treatment sequence in this population of mRCC patients was related to OS, with a higher observed OS in patient receiving axitinib as a second line of treatment, as compared with second line sunitinib or everolimus, after controlling for important confounders including the patient performance status.[31] Such results should be confirmed by analyzing the exhaustive database, since we observed in our case study sample only 4 subjects receiving axitinib.

This case study has several limitations. First, the small sample size associated with the use of the sampled database precluded analyses of the effect of the year of TKI initiation on healthcare use behavior and treatment sequences (given successive changes in the mRCC guidelines). In this context, a higher level of granularity in the description of care (i.e. type of specialist or hospital practitioner) or the addition of other variables corresponding to access to other types of healthcare professionals (i.e. nurses, social workers) would provide a more in-depth description of the three clusters. Caution is required in the generalization of these results to the entire French population of mRCC patients.

Second, reimbursement data did not exhaustively capture the care pathways of mRCC patients taking TKI, since cancer subjects can be included in clinical trials in which drugs are provided free-of-charge by pharmaceutical companies, without reimbursement retrievable. This may introduce a classification bias when studying treatment sequences, with subjects being incorrectly considered as having "no oral anticancer treatment"; or in the identification of patterns of suboptimal management. However, since patients enrolled in clinical trials constitute a very specific population, characterized by a strict follow-up with regularly programmed hospital appointments, exclusion of such patients is not expected to introduce a flawed interpretation of patterns of suboptimal ambulatory management, which can be observed in the real-life setting.

Third, despite the use of a sequential pattern mining method, we did not study the time between consecutive cares. This information is important, since short intervals between consecutive cares may indicate disease worsening. More complex methods could be used in this context, such as the Hirate Yamana algorithm, which adds a temporal dimension to pattern mining. Finally, we used the Charlson comorbidity index to assess comorbidity, but more validated measures, not retrievable from the EGB database, such as the Memorial Sloan-Kettering Cancer Center prognostic score[32] or Heng risk score[33] might provide more accurate information about the ability of the CM-SPAM/FAMD method to discriminate between subjects in terms of prognosis.

As a perspective, in the context of growing healthcare expenditures and shrinking health budgets, organizational initiatives based on data-driven interventions might bring opportunities to improve mRCC management. Our methodology, using EGB sample, gave preliminary results which should be reproduced in the exhaustive national database, which would allow a more territorialized analysis, since EGB is not representative at the scale of the region and the department. Analysis on data from the SNIIRAM could help identifying remote territories with specific healthcare resources requirements, and tailoring efficient healthcare interventions.

Conclusion

This case study demonstrates that the use of data from administrative healthcare databases, coupled with sequence mining and clustering methods, can identify homogeneous groups of individuals in terms of prognosis and healthcare use behaviors, facilitating the identification of particular points in healthcare trajectories at which health-promoting actions are required.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethical approval

Access to the database is legally authorized without the need for permission from the national data protection agency (CNIL). The study protocol was approved by an ethics committee (CER REF 2020-02-20).

Data availability

The data supporting the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID iD

Cyril Baudrier https://orcid.org/0000-0002-6404-1560

References

1. Bedke J, Albiges L, Capitanio U, et al. The 2021 Updated European association of urology guidelines on renal cell carcinoma: immune checkpoint inhibitor–based combination therapies for treatment-naive metastatic clear-cell renal cell carcinoma are standard of care. *Eur Urol* 2021, Epub ahead of print 29 May 2021. DOI: 10.1016/j.eururo.2021.04.042
2. George S, Faccone J, Huo S, et al. Real-world treatment patterns and sequencing for metastatic renal cell carcinoma (mRCC): Results from the Flatiron database. *J Clin Oncol* 2021; 39: 286–286.
3. Finek J, Demlova R, Kopeckova K, et al. Treatment sequences in metastatic renal cell carcinoma: Efficacy results from the Czech registry (RENIS). *Ann Oncol* 2018; 29: viii316–viii317.
4. Ambavane A, Yang S, Atkins MB, et al. Clinical and economic outcomes of treatment sequences for intermediate- to poor-risk advanced renal cell carcinoma. *Immunotherapy* 2020; 12: 37–51.
5. Renet S, Maritaz C, Lotz J-P, et al. [Care pathways of cancer patients: Modeling and risks analysis induced by oral anticancer drugs]. *Bull Cancer (Paris)* 2016; 103: 345–352.
6. *Le Plan cancer 2014-2019-Les Plans cancer*, https://www.e-cancer.fr/Institut-national-du-cancer/Strategie-de-lutte-contre-les-cancers-en-France/Les-Plans-cancer/Le-Plan-cancer-2014-2019 (accessed 28 June 2021).
7. *Suivi des patients atteints de cancer : les généralistes favorables à des échanges renforcés avec l'hôpital | Direction de la recherche, des études, de l'évaluation et des statistiques*, https://drees.solidarites-sante.gouv.fr/publications/etudes-et-resultats/suivi-des-patients-atteints-de-cancer-les-generalistes-favorables (accessed 12 November 2021).

8. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008; 61: 344–349.
9. Tuppin P, de Roquefeuil L, Weill A, et al. French national health insurance information system and the permanent beneficiaries sample. *Rev DÉpidémiologie Santé Publique* 2010; 58: 286–290.
10. SPF. *Croisement de deux bases médico-administratives : méthodologie et étude descriptive pour une application à la surveillance épidémiologique des cancers*. Seconde étape de l'étude exploratoire du croisement PMSI-ALD 2006-2008, https://www.santepubliquefrance.fr/notices/croisement-de-deux-bases-medico-administratives-methodologie-et-etude-descriptive-pour-une-application-a-la-surveillance-epidemiologique-des-canc (accessed 28 June 2021).
11. Caisse Nationale d'Assurance Maladie (CNAM). *Méthodologie médicale de la cartographie des pathologies et des dépenses*. Paris, France: Caisse Nationale d'Assurance Maladie, 2020. version G7 (années 2012 à 2018), https://www.ameli.fr/fileadmin/user_upload/documents/Methodologie_medicale_cartographie.pdf (accessed 25 March 2021).
12. Quan H, Li B, Couris CM, et al. Updating and Validating the Charlson Comorbidity Index and Score for Risk Adjustment in Hospital Discharge Abstracts Using Data From 6 Countries. *Am J Epidemiol* 2011; 173: 676–682.
13. Bice TW and Boxerman SB. A Quantitative Measure of Continuity of Care. *Med Care* 1977; 15: 347–349.
14. *Indicateurs d'observance aux traitements à partir des données du SNDS*. Nykøbing: HEVA, 2020, https://hevaweb.com/fr/articles/indicateurs-dobservance-aux-traitements-a-partir-des-donnees-du-snds/101 (accessed 28 June 2021).
15. Gomariz A, Campos M and Thomas R. Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information. In: 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014.
16. Pagès J. Analyse factorielle de données mixtes. *Rev Stat Appliquée* 2004; 52: 93–111.
17. *HCPC-Hierarchical Clustering on Principal Components: Essentials-Articles-STHDA*, http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/117-hcpc-hierarchical-clustering-on-principal-components-essentials/(accessed 28 June 2021).
18. *ALD n°30-Cancer du rein de l'adulte. Haute Autorité de Santé*, https://www.has-sante.fr/jcms/c_985455/fr/ald-n30-cancer-du-rein-de-l-adulte (accessed 25 July 2021).
19. Jay N, Nuemi G, Gadreau M, et al. 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.
20. Le Meur N, Gao F and Bayat S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Serv Res* 2015; 15: 200.
21. Pinaire J, Chabert E, Azé J, et al. Sequential pattern mining to predict medical in-hospital mortality from administrative data: application to acute coronary syndrome. *J Healthc Eng* 2021; 2021: 1–12.
22. Sayadi S, Geffard E, Südholt M, et al. Secure Distribution of Factor Analysis of Mixed Data (FAMD) and its application to personalized medicine of transplanted patients. In: Woungang I and Enokido T (eds). *Advanced Information Networking and Applications*. Cham: Springer International Publishing, pp. 507–518.
23. Wang C, Kuo H-C, Cheng S-F, et al. Continuity of care and multiple chronic conditions impact frequent use of outpatient services. *Health Informatics Journal* 2020; 26: 318–327. https://journals.sagepub.com/doi/10.1177/1460458218824720?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed (accessed 24 February 2022).
24. *Cancers du rein: vivre avec et après la maladie*. Villejuif, France: Fondation ARC pour la recherche sur le cancer, https://www.fondation-arc.org/cancer/cancer-rein/suivi-apres-cancer (accessed 14 November 2021).

25. Ben-Assuli O, Padman R and Shabtai I. Exploring trajectories of emergency department visits using a laboratory-based indicator of serious illness. *Health Informatics J* 2020; 26: 205–217.
26. Dowd B, Karmarker M, Swenson T, et al. Emergency department utilization as a measure of physician performance. *Am J Med Qual Off J Am Coll Med Qual* 2014; 29: 135–143.
27. Agarwal M, Thareja N, Benjamin M, et al. Tyrosine Kinase Inhibitor-Induced Hypertension. *Curr Oncol Rep* 2018; 20: 65.
28. Wolter P, Hendrickx T, Renard V, et al. Adherence to oral anticancer drugs (OAD) in patients (pts) with metastatic renal cancer (mRCC): First results of the prospective observational multicenter IPSOC study (Investigating Patient Satisfaction with Oral Anti-cancer Treatment). *Journal of Clinical Oncology*; 30: 4622–4622 https://ascopubs.org/doi/10.1200/jco.2012.30.15_suppl.4622 (accessed 24 February 2022).
29. Slejko JF, Rueda J-D, Trovato JA, et al. A comprehensive review of methods to measure oral oncolytic dose intensity using retrospective data. *J Manag Care Spec Pharm* 2019; 25: 1125–1132.
30. *Présentation du projet*. Philadelphia, Pennsylvania: Oncolink, https://therapiesorales-onco-link.fr/presentation-du-projet/(accessed 24 February 2022).
31. Finek J, Demlova R, Kopeckova K, et al. Treatment sequences in metastatic renal cell carcinoma: Efficacy results from the Czech registry (RENIS. *Ann Oncol* 2018; 29: 316–317.
32. Fiala O, Finek J, Poprach A, et al. Outcomes According to MSKCC Risk Score with focus on the intermediate-risk group in metastatic renal cell carcinoma patients treated with first-line sunitinib: a retrospective analysis of 2390 Patients. *Cancers* 2020; 12: 808.
33. Ko JJ, Xie W, Kroeger N, et al. The International Metastatic Renal Cell Carcinoma Database Consortium model as a prognostic tool in patients with metastatic renal cell carcinoma previously treated with first-line targeted therapy: a population-based study. *Lancet Oncol* 2015; 16: 293–300.