AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

Exploring long-term breast cancer survivors' care trajectories using dynamic time warping-based unsupervised clustering

Alexia Giannoula, PhD*,[1,2,3], **Mercè Comas, PhD**[1,3], **Xavier Castells, PhD**[1,3],
Francisco Estupiñán-Romero, MS[3,4], **Enrique Bernal-Delgado** (iD)**, PhD**[3,4], **Ferran Sanz, PhD**[1,2],
Maria Sala, PhD*,[1,3]

[1]Epidemiology and Evaluation Department, Hospital del Mar Research Institute (IMIM), Barcelona, 08003, Spain, [2]Research Programme on Biomedical Informatics (GRIB), Department of Medicine and Life Sciences (MELIS), Hospital del Mar Research Institute (IMIM), Universitat Pompeu Fabra, Barcelona, Spain, [3]RICAPPS Red de Investigación en Cronicidad, Atención Primaria Y Promoción de la Salud, Spain, [4]Data Science for Health Services and Policy Research Group, Institute for Health Sciences (IACS), Zaragoza, Aragon, 50009, Spain

*Corresponding authors: Alexia Giannoula, PhD, Epidemiology and Evaluation Department, Hospital del Mar Research Institute (IMIM), Calle Aiguader 88, Barcelona, 08003, Spain (alexia.giannoula@upf.edu); Maria Sala, PhD, Epidemiology and Evaluation Department, Hospital del Mar Research Institute (IMIM), Calle Aiguader 88, Barcelona, 08003, Spain (msalaserra@psmar.cat)

Abstract

Objectives: Long-term breast cancer survivors (BCS) constitute a complex group of patients, whose number is estimated to continue rising, such that, a dedicated long-term clinical follow-up is necessary.

Materials and Methods: A dynamic time warping-based unsupervised clustering methodology is presented in this article for the identification of temporal patterns in the care trajectories of 6214 female BCS of a large longitudinal retrospective cohort of Spain. The extracted care-transition patterns are graphically represented using directed network diagrams with aggregated patient and time information. A control group consisting of 12 412 females without breast cancer is also used for comparison.

Results: The use of radiology and hospital admission are explored as patterns of special interest. In the generated networks, a more intense and complex use of certain healthcare services (eg, radiology, outpatient care, hospital admission) is shown and quantified for the BCS. Higher mortality rates and numbers of comorbidities are observed in various transitions and compared with non-breast cancer. It is also demonstrated how a wealth of patient and time information can be revealed from individual service transitions.

Discussion: The presented methodology permits the identification and descriptive visualization of temporal patterns of the usage of healthcare services by the BCS, that otherwise would remain hidden in the trajectories.

Conclusion: The results could provide the basis for better understanding the BCS' circulation through the health system, with a view to more efficiently predicting their forthcoming needs and thus designing more effective personalized survivorship care plans.

Key words: data mining; longitudinal analysis; breast cancer survivors; healthcare services.

Background and significance

Long-term breast cancer survivors (BCS) are those patients who survive free from cancer recurrence or the appearance of a new primary cancer for at least 5 years after their diagnosis.[1] In the past decades, the breast cancer landscape has experienced significant changes, manifested by a remarkable rise of both its incidence rate, due mainly, to advances in the public-health screening programs and also, by a steady decline in the overall breast cancer mortality, attributed to earlier detection and more effective treatment options.[2–8] As a consequence, the number of BCS has notably increased, thereby rendering the medical challenges associated with their long-term survivorship a major concern, as these women are at high risk of suffering from late and long-term effects of their treatments, subsequent primary cancers, pre-existing comorbidities, and emotional distress. Additionally, the interplay with aging and other chronic diseases further increases the complexity.[8–11]

As the prevalence of BCS increases, there is a growing need to address gaps in cancer survivorship resources.[12–17] Up to date, only a limited number of works has investigated the usage of healthcare services by BCS, where the focus of attention has been, mainly, on studying their adherence to the guidelines for surveillance,[18–23] assessing their risk of recurrence,[24,25] or describing their long-term multimorbidity.[26] Among the limitations of these works are either the omission of the temporal dimension and/or small sample size, as well as the use of descriptive or other statistical analysis methods, such as descriptive univariate analysis or multivariate linear/logistic regression models, known to be restrictive with respect to the range and depth of useful information that can be extracted.

At the same time, the massive digitization of healthcare in the recent years, through clinical big data and the exponential growth of electronic health records (EHRs), has promoted

the application of data mining in the field of biomedical research.[27–32] Data-mining technology has shown great potential in harnessing a wealth of knowledge from large volumes of medical data that otherwise would remain hidden in the patients' clinical histories.

In this article, an innovative methodological framework employs a combination of unsupervised clustering, time signal processing, visualization, and network analysis in order to describe the long-term BCS' care trajectories. As care trajectories, we define a finite time sequence of chronologically ordered registries of healthcare services used by an individual. To this end, one of the longest existing retrospective longitudinal cohorts is employed (SURvival Breast CANcer [SURBCAN], Spain), with 6214 female BCS and a time span of up to 16 years after their cancer diagnosis. The clustering methodology for disease trajectories presented in the study of Giannoula et al[33] was adapted and expanded, by elaborating a new distance metric appropriate for the type of health data under consideration (healthcare service-use codes). Furthermore, a novel visualization scheme is implemented with the aid of the Cytoscape bioinformatics tool,[34] typically used for visualizing molecular interaction networks and biological pathways. The tool transforms the identified care patterns to an approximated directed and weighted network diagram of frequently visited services, in order to describe these women's transitions through the healthcare system. A wealth of aggregated patient and time information is, at the same time, revealed.

To the best of our knowledge, this is the first application of data mining on BCS, with the goal of identifying temporal patterns (clusters) of healthcare-service transitions over a 5-year follow-up (subsequent to their initial 5-year survival period). The analysis is also performed on a group of 12 412 women without history of breast cancer (non-breast cancer [NBC]) for comparison purposes. Two use cases of particular interest are investigated: (1) radiology, since all guidelines recommend an annual medical consultation and mammography for active surveillance in BCS. Furthermore, depending on their risk of recurrence, they may undergo other diagnostic and surveillance tests for breast cancer-associated diseases and their treatment; and (2) hospital admission, as it may represent an indicator of the patients' severity or healthcare burden.[35] Specifically, although the risk of recurrence and adverse effects are higher in the first few years after diagnosis, secondary breast cancer events can occur at any time, such that, long-term BCS continue to experience health problems and disruptions even decades after diagnosis.

The findings of this study are expected to serve as a preliminary basis for better understanding the BCS' needs and evolution, with a view to developing more effective personalized survivorship care plans that seek to improve their quality of life, further increase survivorship and eventually, improve resource management. The proposed methodology can be readily expanded in order to incorporate additional time-dependent health data of the BCS and in this sense, be applied to a wide range of longitudinal cohort studies in biomedicine.

Methods
Dataset
The longitudinal SURBCAN cohort, in the framework of the Research Network on Health Services in Chronic Diseases (REDISSEC),[36] is an observational population-based retrospective cohort that includes long-term BCS and non-breast cancer (NBC) from 5 Spanish regions: Catalonia (Hospital

del Mar and Information System for the Development of Research in Primary Care [SIDIAP]), Andalucia (Hospital Costa del Sol), Madrid (Hospital 12 de Octubre and Primary Care), Navarra (entire region), and Aragon (EpiChron Cohort, entire region). The cohort data has been collected from sources of primary and hospital care, as well as tumor registries and includes all the individuals' contacts with the Spanish national healthcare services (primary, specialty and hospital care, including emergency visits, hospital admissions, and diagnostic tests), their sociodemographic and lifestyle information (eg, age, nationality, health coverage, smoking and alcohol consumption), pharmaceutical prescriptions, comorbidities, and cancer information (such as size, histological subtype, differentiation grade, metastatic state, estrogen/progesterone receptor status, among others).

For the objectives of this study, 6214 long-term female BCS and 12 412 female NBC from the SURBCAN dataset are used. The BCS (>18 years old) had their breast cancer diagnosis between 2000 and 2006, survived the initial 5-year period and were monitored for 5 years starting on January 1, 2012, until December 31, 2016. In order to be included in the BCS group, they were required to be alive at the beginning of the follow-up period (January 1, 2012) and to have attended at least once the primary-care health services of the country within that period (2012-2016).[19,22] For example, the BCS who were diagnosed in 2000 had survived 11 years before the follow-up started (January 1, 2012), reaching in the end (December 31, 2016) a maximum survival time of 16 years. The NBC control group was formed based on primary-care registries and consists of selected women without a breast cancer diagnosis, matched by age and administrative health area with the BCS. For both the BCS and NBC, their contacts with the healthcare national services during the follow-up are extracted (visit codes, see below) in order to be analyzed in this work.

Unsupervised clustering of the shared care trajectories
A schematic illustration of the workflow of the proposed methodology is presented in Figure 1. Care trajectory is defined to be a finite time sequence of chronologically ordered healthcare services used by an individual[33,37,38] (this definition should be distinguished from the end-of-life illness trajectories of Murray et al[39] and Cohen-Mansfield et al[40]) Therefore, assuming a longitudinal cohort of N long-term female BCS, the care trajectory p of a patient is composed of K sequential numerical visit codes, described in Table 1, that is,

$$p = \{p_1, p_2, \ldots, p_K\}, \qquad (1)$$

where p_k denotes a healthcare service visited by the patient at a discrete time instant t_k ($k = 1, 2, \ldots, K$). Subsequently, a pairwise comparison of the extracted care trajectories of all patients is performed in order to find sub-sequences of distinct healthcare visit codes that are shared in the exact same order by 2 or more individuals (repetitions are ignored). In this manner, a list of shared care trajectories is identified of different lengths L (number of visited services, $L \geq 2$).

These trajectories need next to be clustered in order to identify groups (clusters) of care patterns that share the same time characteristics. For this reason, a significant adaptation and expansion of the unsupervised clustering technique of Giannoula et al[33] for disease trajectories is performed. The dynamic time warping (DTW) algorithm is used to compare the
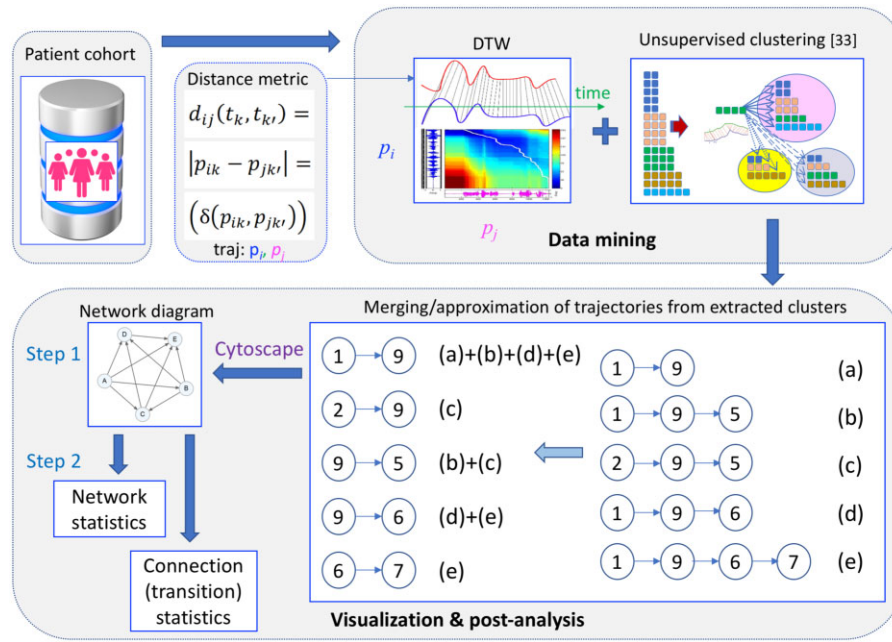
Figure 1. Schematic representation of the workflow of the proposed methodology, illustrating its basic components: patient cohort (BCS/NBC), elaboration of the distance metric for dynamic time warping (DTW), data mining (DTW and unsupervised clustering algorithm), visualization (step 1: merging of trajectories, network generation with Cytoscape), post-analysis (step 2: network and transition statistics). Description of the visit codes for the example of trajectory merging (a)-(e) can be found in Table 1.

Table 1. Description (full and abbreviation) of the healthcare services that correspond to each of the 11 numerical visit codes (1-11) used.

Code	Service	Abbreviation
1	Family doctor	Fam doc
2	Nursing primary care	Nurse
3	Other primary care	Other PC
4	Emergency primary care	ER PC
5	Outpatient care	Outpatient
6	Hospital emergency	ER Hosp
7	Hospital admission	Admission
8	Laboratory	Lab
9	Radiology	Radiol
10	Rehabilitation	Rehab
11	Psychology	Psych

aforementioned trajectories of variable lengths, durations and time intervals, in order to identify shared temporal patterns that may be hidden into the trajectories.[41,42]

Initially, an appropriate distance metric is considered for use in the DTW (Figure 1: distance metric). Specifically, the distance between 2 shared care trajectories $p_i = \{p_{i1}, p_{i2}, \ldots, p_{L_i}\}$ and $p_j = \{p_{j1}, p_{j2}, \ldots, p_{L_j}\}$ of lengths L_i, L_j, correspondingly, is translated into comparing its respective individual visit codes p_{ik} $(k = 1, 2, \ldots, L_i)$ and $p_{jk'}$ $(k' = 1, 2, \ldots, L_j)$. These are represented by categorical variables, as described in Table 1. Therefore, a delta (δ) function is proposed as distance metric for the DTW algorithm, that is,

$$d_{ij}(t_k, t_{k'}) = |p_{ik} - p_{jk'}| = (\delta(p_{ik}, p_{jk'})) = \begin{cases} 1, & \text{if } p_{ik} = p_{jk'} \\ 0, & \text{if } p_{ik} \neq p_{jk'} \end{cases}$$

$$(2)$$

The time instants t_k and $t_{k'}$ can be distinct in 2 patients, since the DTW permits variability in time scaling, length and in-between intervals. Finding similarities between the extracted shared care trajectories corresponds, in principle, to finding exact matches within the compared sequences based on (2). The final *accumulated distance* or *global cost*[41] between the 2 trajectories is finally calculated and introduced into the unsupervised clustering algorithm described below.

Each trajectory from the list of shared care trajectories is compared to those already assigned to different clusters, based on the DTW algorithm and the calculated distance of (2). Then, according to a predetermined user-defined threshold that aims at adjusting the clustering granularity, the assessed trajectory is either assigned to the cluster with which it shares the most similarities (ie, lowest averaged distance) or it is left unassigned and forms a new cluster by itself if the calculated mean distance is lower than the aforementioned threshold.[33] The algorithm terminates when there are no other shared care trajectories to be considered (Figure 1: DTW-based unsupervised clustering). Each extracted cluster is composed of a number of trajectories that reflect a principal care pattern with one or more variations (intermediate visiting service(s) distinct from those of the pattern).

Post-processing analysis and visualization of the identified clusters

For the analysis of the extracted clusters, a novel 2-step post-processing framework has been developed (Figure 1: visualization and post-analysis). Initially, the 3 most frequent trajectories of all clusters associated with an investigated pattern (eg, that of radiology) are collected and merged together into a condensed directed and weighted network diagram of connected services, such that each pairwise connection (service-to-service transition) appears only once (see example in Figure 1). The generated network is visually represented using the open-source bioinformatics software platform Cytoscape 3.9.1[34] and permits acquiring a global view of the

patients' transitions through the different services (step 1). Subsequently, summarized global network statistics are calculated: these include the patients' mean age, mortality rate, mean/median number of comorbidities at the start of the follow-up, mean number of visits to healthcare services and median time intervals (step 2: network statistics in Figure 1).

It also shows how individual network connections can be further explored in order to bring out additional useful information (time/patient-related) about specific service-to-service transitions (step 2: connection statistics in Figure 1). This is achieved by considering each time the expanded subset of patients (after merging trajectories) associated with a specific network connection. Both the global and individual connection statistics are graphically illustrated on the network by adjusting its corresponding node and line properties (eg, color, width, size, etc.).

In order to assess the statistical significance of the differences in the patients' characteristics and time intervals between different groups, the following statistical tests are employed: the 2-sample t-test for the mean age, mean number of comorbidities at the start, and the mean number of visits to a healthcare service, the Pearson's χ^2-test for the mortality rate, and the 2-sided Wilcoxon rank sum test for the median number of comorbidities at start and median time intervals, all at 5% confidence interval (P-values <.05), followed by Bonferroni correction.

Results

Extraction and clustering of the shared care trajectories

Using the proposed methodological pipeline, a total of 5160 shared care trajectories were, initially, extracted from 6214 female BCS for lengths L (number of common distinct visits) varying between 2 and 8. Their distribution with respect to L is described in Table S1. In a similar manner, 7454 trajectories were extracted from 12 412 female NBC, whose respective length distribution is shown in the same table. The 10 most frequent shared trajectories identified for both the case and control group are shown in Tables S2 and S3, respectively, for lengths $L = 2$-7. The trajectories corresponding to $L = 8$ are not shown due to space reasons.

Subsequently, the proposed DTW-based unsupervised clustering algorithm was applied to the extracted care trajectories of each investigated group (5160 in total for the BCS and 7532 for the NBC as mentioned above). In order to filter out trajectories shared by a relatively low number of patients, a cut-off value of 60 and 120 was set as the minimum number of individuals per trajectory for BCS and NBC, respectively (corresponding to ∼1% of the total population in each case). This resulted in a total number of 251 (226) clusters extracted for the BCS (NBC) group, the distribution of which, is shown in Figure S1A and 1B for a given number of assigned trajectories. An example of how an extracted cluster for the BCS group looks like is reported in Table S4.

Post-processing analysis and visualization of the extracted clusters

Two cases of interest are thoroughly assessed hereafter, that is, the patterns associated with the use of the service of radiology and hospital admission. Results obtained from all extracted clusters (251 and 226 for BCS and NBC, respectively) are provided for illustrative purposes as they represent almost the totality of the original dataset (98.1% and 96.6%, respectively). With reference to these, the corresponding trajectory networks are shown in Figure 2A and B and aggregate data is reported in the far right of Tables 2 and 3. The displayed networks encompassed a total of 567 (569) trajectories in BCS (NBC) and contained 74 (75) total connections, demonstrating a complex overall use of the services realized by either group (as measured by the total number of transitions). Furthermore, darker colors in the majority of the connecting lines and target arrows in the BCS network, indicate higher mortality rates and number of comorbidities at the start of the follow-up for most service-to-service transitions than the NBC group.

Pattern of the radiology-service use

In the pattern of the radiology-service use, 51 clusters in total were retrieved for the BCS group, in which the identified pattern is characterized by the presence of the service of radiology (code 9 in Table 1) in all trajectories of the above clusters, in any order. This resulted in a total of 139 shared care trajectories and 2689 patients (43.3% of the entire BCS) associated with the radiology pattern. In a similar manner, 46 clusters consisting of 140 trajectories and represented by 4200 individuals (33.8% of the entire NBC) were retrieved for the respective NBC radiology pattern.

The corresponding trajectory networks are shown in Figure 3A and B for either group. A comparison between the 2 reveals that, overall, the BCS assigned to radiology clusters exhibited higher mortality rates and a higher number of comorbidities at the start of the follow-up period compared to the NBC group, as reflected by the darker colors of the respective target arrows and connecting lines for the majority of the individual connecting paths. This is even more pronounced for those patients sharing the connections between radiology and hospital emergency or hospital admission services.

Comparative statistics referring to the radiology patterns of both BCS and NBC can be found in Table 2. Statistically significant differences between the BCS and NBC groups are observed in mortality rate (10.6% vs 6.7%, P-value <.001) and number of comorbidities at the start of the follow-up (3.9/3 vs 2.9/2, P-value <.001). Summarized information about their usage of healthcare services is shown in Table 3, as resulted by calculating the mean (and standard deviation) of the total number of visits that women from the corresponding group made to each healthcare service. As can be observed, the BCS visited significantly more often outpatient care (22.9 vs 13.9 mean visits, P-value <.001), radiology (11.2 vs 6.7 mean visits, P-value <.001), and hospital admission (0.9 vs 0.7 mean visits, P-value <.001) than the equivalent NBC group. Interestingly, they visited the laboratory less often than the NBC (4.4 vs 5.6, P-value <.001), as well as emergency primary care (1.8 vs 2.1, P-value = .002). No statistically significant differences were found for the rest of the services used.

Despite the above notable differences between the BCS and NBC radiology patterns, individuals from either group used most of the healthcare services to a comparable extent and presented similar comparative characteristics as in the case of all extracted clusters (see Tables 2 and 3). On the other side, both groups made a higher usage of outpatient care, hospital emergency, laboratory, and radiology than the all-cluster respective groups, indicating a more intense use of these
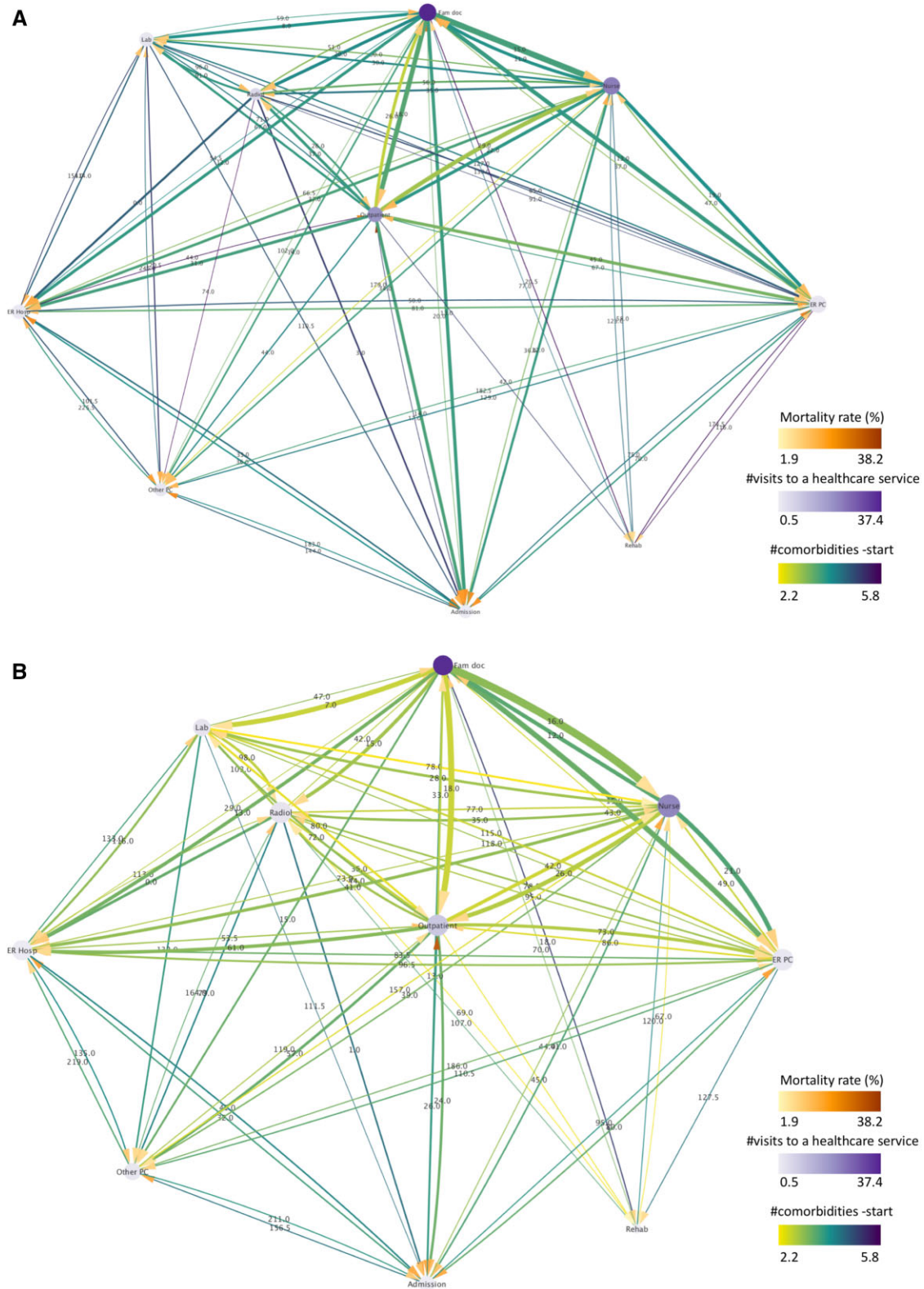
Figure 2. Directed trajectory network generated by all extracted clusters for the (A) BCS and (B) NBC groups. The width of the lines is proportional to the number of patients transiting between 2 services (normalized by the total number of patients of the pattern) and their color is mapped to the mean number of their comorbidities at the start of the follow-up. The size of the nodes is proportional to their degree (number of total connections leading into and exiting from a node) and their color represents the mean number of the patients' visits to a healthcare service. The color of the target arrows denotes mortality rate. The label of the lines illustrates the median time interval between 2 services measured in days. All line and node properties (except for the line width) have been normalized between the BCS and NBC to facilitate their direct comparison.

Table 2. Women's characteristics calculated for the trajectory networks of the breast cancer survivors (BCS) and non-breast cancer (NBC) groups, associated with the patterns of radiology and hospital admission.

	Radiology			Admission			All clusters		
	BCS	NBC	P-value[a]	BCS	NBC	P-value[a]	BCS	NBC	P-value[a]
Number of women, N (%)[b]	2689 (43.3%)	4200 (33.8%)	–	2151 (34.6%)	2202 (17.7%)	–	6097 (98.1%)	11 988 (96.6%)	–
Mean age (SD)	65.3 (12.1)	65.9 (12.0)	.108	68.5 (13.2)	70.5 (12.3)	<.001	66.7 (12.5)	66.7 (12.5)	.963
Mortality, N_{exitus} (%)[c]	284 (10.6%)	280 (6.7%)	<.001	466 (21.7%)	352 (16.0%)	<.001	755 (12.4%)	915 (7.7%)	<.001
Comorbidities at start[d] Mean (SD)	3.9 (3.9)	2.9 (3.7)	<.001	3.8 (4.0)	3.7 (3.9)	.697	3.7 (3.8)	3.2 (3.4)	<.001
Median [Q1, Q3]	3 [1, 5]	2 [0, 4]	<.001	3 [1, 5]	3 [1, 5]	0.701	3 [1, 5]	2 [1, 5]	<.001

The same data are finally listed in the case of all extracted clusters.
  [a] P-values assess statistically significant differences (<.05) between female BCS and NBC. The 2-sample *t*-test was used for the mean age and mean number of comorbidities at start, Pearson's χ^2-test for the mortality rate and the 2-sided Wilcoxon rank sum test for the median number of comorbidities at start.
  [b] Number *N* of female BCS within a group of clusters (% of the total BCS).
  [c] Number N_{exitus} of women with confirmed deaths (% of those sharing the transition).
  [d] Number of comorbidities at the start of the follow-up, mean (standard deviation) and median (first, third quartiles [Q1, Q3]).

Table 3. Number of visits (mean and standard deviation) to each healthcare service for the breast cancer survivors (BCS) and non-breast cancer (NBC) groups, associated with the patterns of radiology and hospital admission.

	Radiology			Admission			All clusters		
	BCS	NBC	P-value[a]	BCS	NBC	P-value[a]	BCS	NBC	P-value[a]
Fam doc (1)	39.5 (30.0)	41.0 (28.2)	.039	46.1 (37.3)	49.4 (31.1)	.001	37.4 (30.6)	36.3 (27.0)	.024
Nurse (2)	24.3 (34.4)	24.3 (29.5)	.958	30.4 (38.1)	34.8 (38.4)	<.001	23.4 (32.8)	21.8 (30.1)	.001
Other PC (3)	2.0 (5.8)	2.0 (5.4)	.610	2.5 (6.5)	3.1 (7.1)	.002	1.8 (5.5)	2.0 (5.9)	.089
ER PC (4)	1.8 (4.0)	2.1 (4.2)	.002	3.0 (6.3)	3.1 (6.5)	.860	1.9 (4.5)	1.9 (4.6)	.210
Outpatient (5)	22.9 (30.8)	13.5 (21.0)	<.001	35.1 (44.7)	21.3 (27.3)	<.001	19.7 (31.2)	8.7 (15.7)	<.001
ER Hosp (6)	1.9 (2.9)	1.9 (2.5)	.705	3.1 (4.9)	2.6 (2.8)	<.001	1.4 (3.3)	0.9 (1.9)	<.001
Admission (7)	0.9 (1.6)	0.7 (1.4)	<.001	2.4 (3.2)	2.1 (2.1)	<.001	0.9 (2.2)	0.5 (1.3)	<.001
Lab (8)	4.4 (4.2)	5.6 (4.6)	<.001	2.5 (4.2)	3.9 (5.3)	<.001	2.0 (3.6)	2.4 (4.0)	<.001
Radiol (9)	11.2 (8.4)	6.7 (6.5)	<.001	8.3 (10.8)	6.6 (8.3)	<.001	5.0 (7.9)	2.4 (5.0)	<.001
Rehab (10)	1.2 (3.0)	1.2 (2.7)	.369	0.8 (2.6)	0.8 (2.4)	.829	0.8 (2.8)	0.8 (3.3)	.518
Psych (11)	0.03 (1.0)	0.02 (0.5)	.519	0.03 (0.9)	0.03 (0.6)	.940	0.05 (1.1)	0.04 (0.9)	.593

The same data are also reported if all extracted clusters are considered.
  [a] P-values assess statistically significant differences (<.0045) between BCS and NBC, as resulted based on the 2-sample *t*-test at the 5% significant level, followed by Bonferroni correction.

services by individuals of the radiology pattern independently of their history or absence of breast cancer.

Pattern of the hospital admission service

In the hospital admission pattern, trajectories within all clusters required the visit code 9 in any order (Table 1). Overall, 22 clusters containing a total of 52 shared care trajectories and represented by 2151 patients were retrieved for the BCS group (34.6%). In a similar manner, 12 clusters comprising 28 shared trajectories of 2202 NBC individuals were retrieved for the NBC group (17.7%). It is of note that the proportion of BCS patients corresponds to approximately twice that of the NBC counterpart, indicating a more intense usage of this service by the former. The corresponding directed trajectory networks, illustrating the individuals' transitions through the identified services, are shown in Figure 4A and B. In general, higher activity can be observed for the network of the BCS compared with that of the NBC, with 33 versus 26 total transitions, respectively (a particularly higher activity is seen for the nursing, emergency primary-care, and outpatient-care nodes). Furthermore, the

BCS of the admission pattern (younger on average than NBC, ie, 68.5 vs 70.5 years old) presented higher mortality rates than their NBC counterpart, similar to the respective radiology patterns, indicated by the darker target arrows in the majority of the paths. Summarized information associated with the admission network can be found in Table 2 (middle part), where a mortality rate of 21.7% was calculated for the BCS group as opposed to 16.0% for the NBC. Interestingly, the corresponding mean/median numbers of comorbidities at the start of the follow-up were comparable. Also, individuals of both BCS/NBC admission groups were older than the radiology-pattern equivalents (65.3/65.9 vs 68.5/70.5 years old, respectively).

Mean values of the total number of contacts of these women with the healthcare system are shown in Table 3. It can be observed that the BCS admission group visited several services significantly more often than the NBC equivalent, that is, not only outpatient care (35.1 vs 21.3, P-value <.001), radiology (8.3 vs 6.6, P-value <.001) and hospital admission (2.4 vs 2.1, P-value <.001) as in the radiology pattern, but also the hospital-emergency service (3.1 vs 2.6,
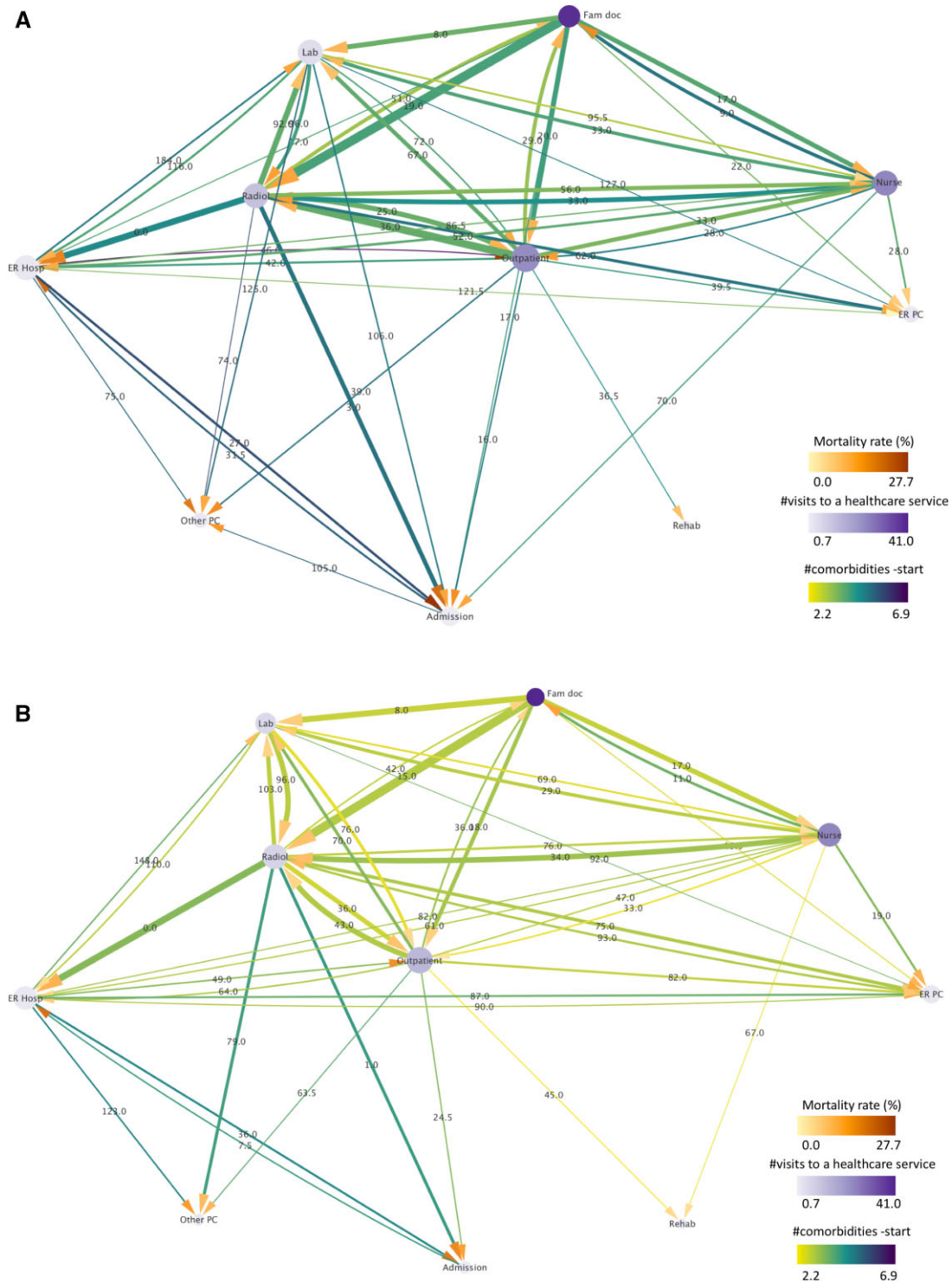
Figure 3. Directed trajectory network of the pattern of radiology for the (A) BCS and (B) NBC groups. The width of the lines is proportional to the number of patients transiting between 2 services (normalized by the total number of patients of the pattern) and their color is mapped to the mean number of their comorbidities at the start of the follow-up. The size of the nodes is proportional to their degree (number of total connections leading into and exiting from a node) and their color represents the mean number of the patients' visits to a healthcare service. The color of the target arrows denotes mortality rate. The label of the lines illustrates the median time interval between 2 services measured in days. All line and node properties (except for the line width) have been normalized between the BCS and NBC to facilitate their direct comparison.

P-value <.001). On the contrary, they made a significantly lower use of the family doctor (46.1 vs 49.4, *P*-value = .001), nursing (30.4 vs 34.8, *P*-value <.001), laboratory (2.5 vs 3.9, *P*-value <.001) and other primary care (3.0 vs 3.1, *P*-value = .002) services.

Apart from the above differences between BCS and NBC, more intense use of the majority of the healthcare services (except for rehabilitation and psychology) was observed for both groups of the hospital-admission pattern when compared with the respective groups of all extracted patterns (see
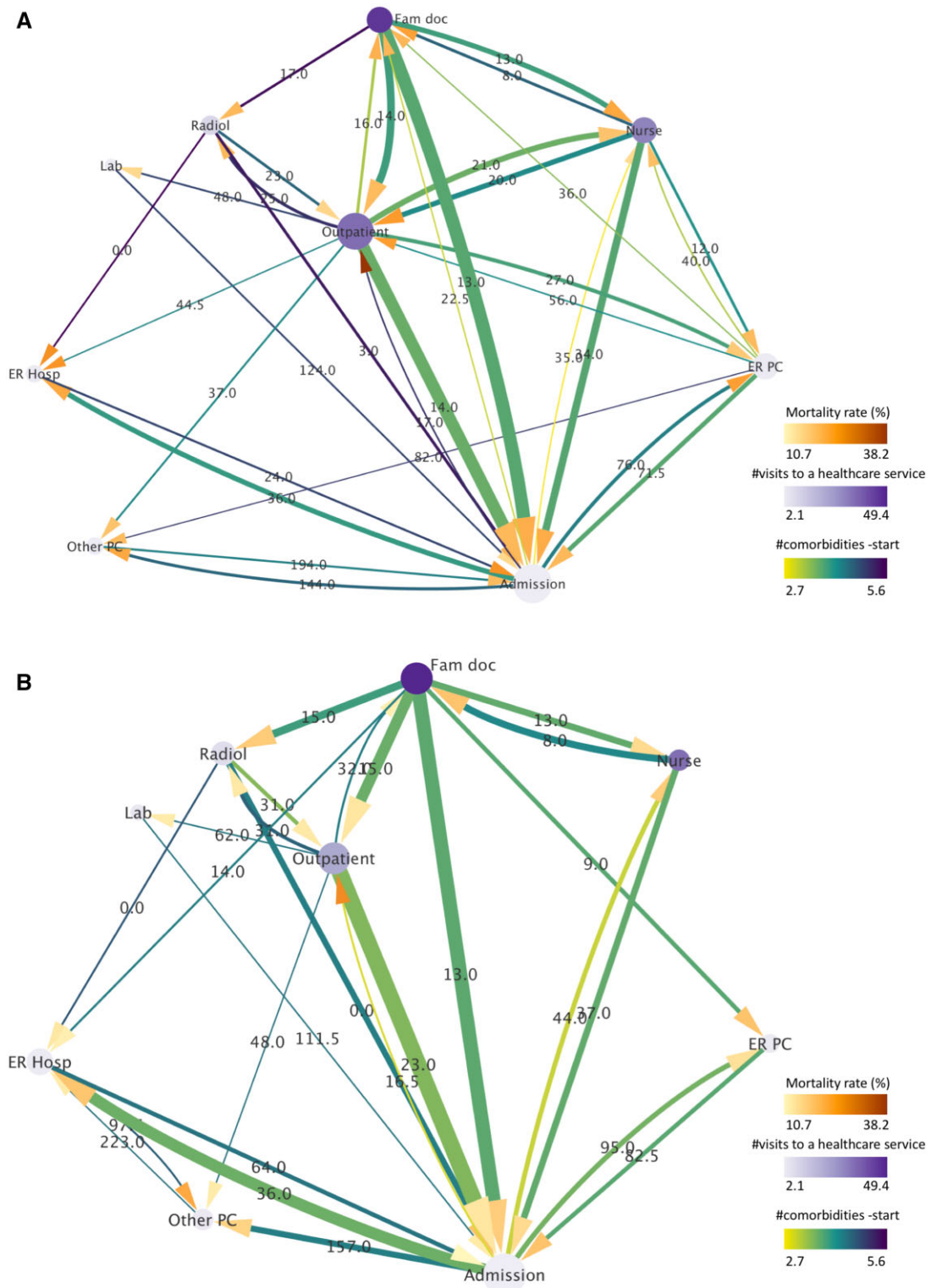
Figure 4. Directed trajectory network of the pattern of hospital admission for the (A) BCS and (B) NBC groups. The width of the lines is proportional to the number of patients transiting between 2 services (normalized by the total number of patients of the pattern) and their color is mapped to the mean number of their comorbidities at the start of the follow-up. The size of the nodes is proportional to their degree (number of total connections leading into and exiting from a node) and their color represents the mean number of the patients' visits to a healthcare service. The color of the target arrows denotes mortality rate. The label of the lines illustrates the median time interval between 2 services measured in days. All line and node properties (except for the line width) have been normalized between the BCS and NBC to facilitate their direct comparison.

Table 3). Finally, when the 2 investigated patterns are compared (Table 3), it is observed that although the admission-pattern individuals (BCS/NBC) visited almost all services more often than those in all clusters (approximately the entire dataset), the radiology-pattern individuals did indeed make even more frequent use of the services of radiology and laboratory.

In-depth analysis of individual service-to-service connections

In order to retrieve additional useful information contained in the identified clusters, the proposed methodology permits further analyzing the patients' transitions through the distinct healthcare services of each pattern. Presenting a detailed description of all network connections is out of the scope of this article. For this reason, only 2 pairwise service-to-service transitions were selected for each pattern (Figures 3 and 4). Aggregated women characteristics and in-between time intervals were calculated for the specific node connections for BCS and NBC and are reported in Table 4. In the same manner, longer and more complex transition trajectories (eg, a series of service transitions) can be readily analyzed.

In the pattern of radiology, the specific transitions from family doctor and outpatient care towards the service of radiology are assessed, as they represent 2 of the most populated ones within the network. It can be observed that a significantly larger proportion of the BCS transited from outpatient to radiology than the NBC (23.5% vs 15%, respectively, when normalized with the total number of patients of the respective BCS/NBC group). Furthermore, individuals of either group transited faster in *family doctor → radiology* than in *outpatient care → radiology* (19/15 vs 36/43 days, respectively). In the first transition, slightly higher mortality rates were found for both BCS/NBC (12.0% and 7.5%, respectively) than those of *outpatient care → radiology* (10.1% and 4.7%).

In the pattern of hospital admission, the transitions from radiology and hospital emergency were selected, as they presented several interesting characteristics that may potentially

require further investigation (see below). The respective median transit times were significantly longer in BCS as compared to NBC when transiting from radiology (3 vs 0 days, respectively) and shorter when transiting from hospital emergency (24 vs 64). Also, the transit times of the first connection (*radiology → admission*) were significantly lower than those of the second connection (*hospital emergency → admission*) for either BCS/NBC (3/0 vs 24/64 days, respectively, for each connection). Furthermore, although the overall mortality rate (admission pattern, Table 2) was calculated to be significantly higher in BCS than NBC (21.7% vs 16.0%, respectively), this did not hold true for the transition of *radiology → admission* (20.6% vs 18.3%, *P*-value = .464). However, for the transition of *hospital emergency → admission*, the difference in the mortality rates was even more pronounced (25.1% vs 10.7%) than that of the entire pattern, indicating a rate of 1 out of 4 BCS that made use of the hospitalization service with confirmed death (one of the highest rates) as opposed to, approximately, 1 out of 10 in NBC (one of the lowest rates). Finally, the mean/median numbers of comorbidities for *radiology → admission* were found to be significantly higher in BCS than in NBC (5.4/4 vs 4.3/3, respectively).

Discussion

An innovative methodological framework employing unsupervised clustering, time-domain signal processing, visualization, and network-analysis techniques was presented in this article for the exploratory analysis of the care trajectories by long-term BCS. A large retrospective longitudinal cohort of Spain was used, containing data for both the BCS and a matched control group (NBC) over a period of 5 years of monitoring (and up to 16 years of survival). An unsupervised clustering technique, adequately adapted from Giannoula et al[33] was applied to the aforementioned trajectories based on a modified DTW algorithm. More than 200 clusters were extracted for both the BCS and NBC, which reflected complex temporal patterns of the use that each group made of the

Table 4. Women characteristics and time information for 2 individual transitions identified in the pattern of (1) radiology (network of Figure 3) and (2) hospital admission (network of Figure 4), for the breast cancer survivors (BCS) and non-breast cancer (NBC) groups.

	Service	Number of women, N (%)[a]	Δt, Median [Q1, Q3][b]	Age, mean (SD)	Mortality, N_exitus (%)[c]	Comorbidities at start[d]	
						Mean (SD)	Median [Q1, Q3]
(1) Family	BCS	1791 (28.8%)	19 [7, 51]	66.6 (12.1)	214 (12.0%)	4.1 (3.9)	3 [2, 6]
Doc→	NBC	3054 (25%)	15 [6, 41]	66.5 (12.3)	229 (7.5%)	3.0 (3.7)	2 [0, 4]
Radiology	*P*-value[e]	–	<.001	.827	<.001	<.001	<.001
(1) Out-	BCS	1462 (23.5%)	36 [12 110]	65.2 (12.0)	147 (10.1%)	3.8 (3.5)	3 [1, 5]
patient→	NBC	1862 (15%)	43 [15 124]	65.9 (11.3)	88 (4.7%)	3.0 (3.5)	2 [0, 5]
Radiology	*P*-value[e]	–	.127	<.001	<.001	<.001	<.001
(2) Radiolo-	BCS	345 (5.6%)	3 [0332]	69.5 (12.5)	71 (20.6%)	5.4 (4.9)	4 [3, 7]
gy→	NBC	497 (4.0%)	0 [0, 24]	71.9 (11.4)	91 (18.3)	4.3 (4.4)	3 [1, 6]
Admission	*P*-value[e]	–	<.001	.015	.464	.002	<.001
(2) ER	BCS	227 (3.7%)	24 [1128]	70.0 (12.4)	57 (25.1%)	4.9 (4.1)	4 [2, 6]
Hosp→	NBC	367 (3.0%)	64 [2253]	71.3 (11.4)	39 (10.7%)	4.5 (3.9)	4 [2, 6]
Admission	*P*-value[e]	–	<.001	.280	<.001	.162	.179

[a] Number N of female BCS sharing a service-to-service transition, N (% of the total BCS).
[b] Time interval between the 2 service visits measured in days: median and first (Q1) and third (Q3) quartiles.
[c] Number N_{exitus} of women with confirmed deaths (% of those sharing the transition).
[d] Number of comorbidities at the start of the follow-up, mean (standard deviation) and median (first, third quartiles).
[e] *P*-values assess statistically significant differences ($<.005$) between female BCS and NBC. The 2-sample *t*-test was used for the mean age and mean number of comorbidities at start, Pearson's χ^2-test for the mortality rate and the 2-sided Wilcoxon rank sum test for the median number of comorbidities at start and median time intervals Δt. Bonferroni correction was also performed.

healthcare services. A novel visualization and post-analysis scheme was next developed by which, directed and weighted trajectory network diagrams of the patients' transitions were generated by means of an approximation process and the Cytoscape tool. Subsequently, aggregated patient and time information resulting from these networks was calculated. This 2-step analysis process of the identified patterns facilitated the direct comparison not only between BCS and NBC, but also, between each other.

On one hand, it was revealed that long-term BCS of either of the 2 investigated patterns did indeed make significantly more intense and complex use of particular healthcare services (eg, radiology, outpatient care, hospital admission) and presented higher mortality rates than the NBC counterpart. BCS of the radiology pattern exhibited, in addition, a higher number of comorbidities at the start of the follow-up. It should be noted that the usage of the radiology service included all kinds of imaging tests, regardless of the pathology or reason for prescription (eg, surveillance, diagnosis, etc.). The differences encountered cannot be necessarily attributed to breast cancer; however, they are indicative of the higher number of healthcare resources used by the BCS and the higher impact on certain negative outcomes compared to the NBC, as mentioned above.

With respect to the pattern of hospital admission, in particular, approximately twice as many individuals were assigned in BCS (despite being relatively younger) than in NBC and the corresponding trajectory network was of higher activity. Furthermore, both groups of the aforementioned pattern used the majority of the services more than the respective groups of all extracted patterns. This is likely attributed to the expected severity associated with individuals who require hospitalization, regardless of a history of breast cancer, and can be also confirmed by the significantly higher mortality rates for both BCS and NBC of the admission pattern as opposed to all clusters (Table 2).

Although similar trends indicating a more intense use of the healthcare resources have been also observed in previous studies,[19,23,43] with the present work, apart from the incorporation of the time factor, a wealth of additional information (patient/time) on service-to-service transitions was revealed and quantified that otherwise would remain undiscovered (eg, specific time information: the BCS transit approximately twice faster in *family doctor → radiology* than in *outpatient care → radiology*, or important mortality information: 1 out of 4 BCS moving from hospital emergency to admission die, as opposed to 1 out of 10 in NBC, see Table 4).

Breast cancer is a heterogeneous disease and, although all guidelines recommend an annual medical consultation and mammography for active surveillance, the frequency and site of appointments among BCS depend on their risk of recurrence and comorbidities.[15,20] In light of this, incorporating specific breast cancer or imaging history for the BCS would be helpful in interpreting the extracted patterns. However, this work aimed at demonstrating the potential of the proposed methodology to extract a wide range of time and patient information hidden into their care trajectories, rather than seeking clinical interpretation of all the identified patterns.

Inherent to the nature of the health dataset used, there may be incomplete data and errors in the visit and date codes, as well as missing or wrong data regarding the patients'

personal information and related clinical history. Furthermore, BCS can transit through the healthcare system not only based on the recommended guidelines and personal health conditions but also according to the organization and available resources of the local health systems, to which they have access. Any limitations in this aspect should be taken into consideration when interpreting the results. With respect to the distance metric used in the DTW, any other metric that permits comparing the care trajectories could be incorporated, such as a weighted metric that may assign specific weights to different healthcare services (if available).

Conclusions

The proposed DTW-based unsupervised clustering technique may lay the groundwork for better predicting the BCS' needs in the long-term, by studying patterns of the usage of healthcare services that they make over time. Although no conclusions on the causality of the extracted patterns can be now drawn, these can set the basis for elucidating specific questions about the BCS' care trajectories, such as, how/why a sub-group of BCS patients transit from one healthcare setting to another/others, what these patients' characteristics are, whether such pathway is adequate or efficient according to the clinical guidelines and outcomes, etc., with the objective of better understanding their long-term needs and improving the healthcare received. For the study of concrete hypotheses, a promising direction for future work could be the exploration of the patients' comorbidities either at the start or during the follow-up.[26,37,38] For example, a specific trajectory associated with higher mortality, may require further investigation in order to verify if the corresponding patients have adhered to the surveillance guidelines or if their comorbidity status requires a more adequate care management.

Similarly, the methodology could be readily expanded in order to incorporate additional longitudinal patient information, such as their prescription drugs, visits to medical specialists, imaging and laboratory data, etc., that among all could facilitate the elaboration of more efficient patient-centered care plans. Overall, this is a flexible unsupervised-clustering methodological framework that can be applied to any longitudinal cohort in biomedicine, whose objective is describing the temporal behavior of a health outcome of interest (eg, disease, condition, event, change in health status, measurement, etc.) by a group of patients.

Javier Salamanca, Óscar Toldos. Grupo de investigación en servicios sanitarios y cronicidad de la Fundación Miguel Servet, Navarra: Javier Baquedano, Rossana Burgui, Javier Gorricho, Berta Ibáñez, Conchi Moreno, Ibai Tamayo.

Author contributions

A. G. designed the methodology, performed the simulations and analysis of data, and drafted the manuscript. M. C. helped with the relevant statistical tests and with interpreting the results. F. E. R., E. B. D., F. S., and X. C. revised the manuscript and provided feedback with respect to epidemiological issues. M. S. supervised the work, helped in interpreting the results, and approved the final manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

Data availability

The data underlying this article cannot be shared publicly for the privacy of individuals that participated in the study (ethics approval CEIM PSMAR 2019/8639/I). The data will be shared on reasonable request to the corresponding author.

Ethics approval

All data contained in the cohort are anonymized and the corresponding ethical approvals (CEIM PSMAR 2019/8639/I) have been granted.

References

1. Cardoso F, Harbeck N, Barrios CH, et al. Research needs in breast cancer. *Ann Oncol*. 2017;28(2):208-217.
2. Giaquinto AN, Sung H, Miller KD, et al. Breast cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(6):524-541.
3. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209-249.
4. Plevritis SK, Munoz D, Kurian AW, et al. Association of screening and treatment with breast cancer mortality by molecular subtype in US women, 2000-2012. *JAMA*. 2018;319(2):154-164.
5. Howlader N, Noone AM, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2018. SEER. 2021. https://seer.cancer.gov/csr/1975_2018/index.html. Accessed March 1, 2023.
6. Dafni U, Tsourti Z, Alatsathianos I. Breast cancer statistics in the European Union: incidence and survival across European countries. *Breast Care (Basel)*. 2019;14(6):344-353.
7. Carioli G, Malvezzi M, Rodriguez T, et al. Trends and predictions to 2020 in breast cancer mortality in Europe. *Breast*. 2017;36:89-95.
8. Arnold M, Morgan E, Rumgay H, et al. Current and future burden of breast cancer: global statistics for 2020 and 2040. *Breast*. 2022;66:15-23.
9. Miller KD, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022. *CA Cancer J Clin*. 2022;72(5):409-436.
10. Bodai BI, Tuso P. Breast cancer survivorship: a comprehensive review of long-term medical issues and lifestyle recommendations. *Perm J*. 2015;19(2):48-79.
11. Carreira H, Williams R, Dempsey H, et al. Quality of life and mental health in breast cancer survivors compared with non-cancer controls: a study of patient-reported outcomes in the United Kingdom. *J Cancer Surviv*. 2021;15(4):564-575.
12. Hewitt ME, Greenfield S, Stovall E, et al., eds. *From Cancer Patient to Cancer Survivor: Lost in Transition*. National Academies Press; 2006.
13. Jacobs LA, Shulman LN. Follow-up care of cancer survivors: challenges and solutions. *Lancet Oncol*. 2017;18(1):e19-e29.
14. Moore HCF. Breast cancer survivorship. *Semin Oncol*. 2020;47(4):222-228.
15. Runowicz CD, Leach CR, Henry NL, et al. American Cancer Society/American Society of clinical oncology breast cancer survivorship care guideline. *CA Cancer J Clin*. 2016;66(1):43-73.
16. González-Castro L, Cal-González VM, Del Fiol G, et al. CASIDE: a data model for interoperable cancer survivorship information based on FHIR. *J Biomed Inform*. 2021;124(13):103953.
17. Petersen C. Patient-generated health data: a pathway to enhanced long-term cancer survivorship. *J Am Med Inform Assoc*. 2016;23(3):456-461.
18. Draeger T, Voelkel V, Schreuder K, et al. Adherence to the Dutch breast cancer guidelines for surveillance in breast cancer survivors: real-world data from a pooled multicenter analysis. *Oncologist*. 2022;27(10):e766-e773.
19. Jansana A, Posso M, Guerrero I, et al. Health care services use among long-term breast cancer survivors: a systematic review. *J Cancer Surviv*. 2019;13(3):477-493.
20. Santiá P, Jansana A, del Cura I, et al.; SURBCAN Group. Adherence of long-term breast cancer survivors to follow-up care guidelines: a study based on real-world data from the SURBCAN cohort. *Breast Cancer Res Treat*. 2022;193(2):455-465.
21. Brauer ER, Long EF, Petersen L, et al. Current practice patterns and gaps in guideline-concordant breast cancer survivorship care. *J Cancer Surviv*. 2021;17(3):906-915.
22. Jansana A, Del Cura I, Prados-Torres A, et al.; SURBCAN group. Use of real-world data to study health services utilisation and comorbidities in long-term breast cancer survivors (the SURBCAN study): study protocol for a longitudinal population-based cohort study. *BMJ Open*. 2020;10(9):e040253.
23. Quyyumi FF, Wright JD, Accordino MK, et al. Factors associated with follow-up care among women with early-stage breast cancer. *J Oncol Pract*. 2019;15(1):e1-e9.
24. Draeger T, Voelkel V, Groothuis-Oudshoorn CGM, et al. Applying risk-based follow-up strategies on the Dutch breast cancer population: consequences for care and costs. *Value Health*. 2020;23(9):1149-1156.
25. Witteveen A, de Munck L, Groothuis-Oudshoorn CGM, et al. Evaluating the age-based recommendations for long-term follow-up in breast cancer. *Oncologist*. 2020;25(9):e1330-e1338.

26. Jansana A, Poblador-Plou B, Gimeno-Miguel A, et al.; SURBCAN Group. Multimorbidity clusters among long-term breast cancer survivors in Spain: results of the SURBCAN study. *Int J Cancer*. 2021;149(10):1755-1767.

27. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6(1):54.

28. Ahmad P, Qamar S, Qasim Afser Rizvi S. Techniques of data mining in healthcare: a review. *Int J Comput Appl*. 2015;120 (15):38-50.

29. Wu W-T, Li Y-J, Feng A-Z, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res*. 2021;8(1):44.

30. Kaur I, Doja MN, Ahmad T. Data mining and machine learning in cancer survival research: an overview and future recommendations. *J Biomed Inform*. 2022;128:104026.

31. Campbell EA, Bass EJ, Masino AJ. Temporal condition pattern mining in large, sparse electronic health record data: a case study in characterizing pediatric asthma. *J Am Med Inform Assoc*. 2020;27(4):558-566.

32. Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J Am Med Inform Assoc*. 2016;23(2):339-348.

33. Giannoula A, Gutierrez-Sacristán A, Bravo Á, et al. Identifying temporal patterns in patient disease trajectories using dynamic time warping: a population-based study. *Sci Rep*. 2018;8 (1):4216-4214. https://doi.org/10.1038/s41598-018-22578-1

34. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504.

35. Maddams J, Utley M, Møller H. A person-time analysis of hospital activity among cancer survivors in England. *Br J Cancer*. 2011;105(Suppl 1):S38-S45. https://doi.org/10.1038/bjc.2011.421

36. REDISSEC—Red de Investigación en Servicios de Salud en Enfermedades Crónicas. Accessed March 3, 2023. https://www.redissec.com/

37. Jensen AB, Moseley PL, Oprea TI, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun*. 2014;5:4022-4010. https://doi.org/10.1038/ncomms5022

38. Giannoula A, Centeno E, Mayer M-A, et al. A system-level analysis of patient disease trajectories based on clinical, phenotypic and molecular similarities. *Bioinformatics*. 2021;37(10):1435-1443.

39. Murray SA, Kendall M, Boyd K, et al. Illness trajectories and palliative care. *BMJ*. 2005;330(7498):1007-1011.

40. Cohen-Mansfield J, Skornick-Bouchbinder M, Brill S. Trajectories of end of life: a systematic review. *J Gerontol B Psychol Sci Soc Sci*. 2018;73(4):564-572.

41. Müller M, ed. Dynamic time warping. In: *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer; 2007:69-84. https://doi.org/10.1007/978-3-540-74048-3_4

42. Bhavani SV, Xiong L, Pius A, et al. Comparison of time series clustering methods for identifying novel subphenotypes of patients with infection. *J Am Med Inform Assoc*. 2023;30(6):1158-1166.

43. Hooning MJ, Aleman BMP, van Rosmalen AJM, et al. Cause-specific mortality in long-term survivors of breast cancer: a 25-year follow-up study. *Int J Radiat Oncol Biol Phys*. 2006;64 (4):1081-1091.