Artificial Intelligence In Medicine 105 (2020) 101855

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Mining post-surgical care processes in breast cancer patients

Lorenzo Chiudinelli[a], Arianna Dagliati[b], Valentina Tibollo[c], Sara Albasini[c], Nophar Geifman[b], Niels Peek[b], John H. Holmes[d], Fabio Corsi[c], Riccardo Bellazzi[a,c], Lucia Sacchi[a,*]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[b] University of Manchester, UK
[c] IRCCS Istituti Clinici Scientifici Maugeri, Pavia, Italy
[d] University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

ARTICLE INFO

Keywords:
Breast cancer
Process Mining
Topic Modelling
Latent Dirichlet Allocation
Temporal Data Analytics
Temporal Electronic Phenotyping
Electronic Health Records

ABSTRACT

In this work we describe the application of a careflow mining algorithm to detect the most frequent patterns of care in a cohort of 3000 breast cancer patients. The applied method relies on longitudinal data extracted from electronic health records, recorded from the first surgical procedure after a breast cancer diagnosis. Careflows are mined from events data recorded for administrative purposes, including procedures from ICD9 – CM billing codes and chemotherapy treatments. Events data have been pre-processed with Topic Modelling to create composite events based on concurrent procedures. The results of the careflow mining algorithm allow the discovery of electronic temporal phenotypes across the studied population. These phenotypes are further characterized on the basis of clinical traits and tumour histopathology, as well as in terms of relapses, metastasis occurrence and 5-year survival rates. Results are highly significant from a clinical perspective, since phenotypes describe well characterized pathology classes, and the careflows are well matched with existing clinical guidelines. The analysis thus facilitates deriving real-world evidence that can inform clinicians as well as hospital decision makers.

1. Introduction

Breast cancer is the most commonly occurring cancer in women. In Italy, about 1 out of 8 women will be diagnosed with breast cancer during her lifetime. While the trend of newly diagnosed cases between 2003 and 2018 has slightly increased (+0,3 %/year), mortality significantly lowered (-0,8 %/year), due to the combination of early screening and better therapeutic options [1]. To improve breast cancer care, beyond randomized clinical trials, the analysis of different diagnostic and treatment patterns on large cohorts is increasingly performed in order to derive guidelines based on higher levels of evidence.

In cancer management, each diagnostic or treatment guideline is associated with a well-defined set of procedures and drug therapies [2]. The sequence of such procedures performed on specific patients can be extracted from the combination of administrative and clinical data routinely collected and stored at the hospital. In this context, it is of interest to investigate if patients can be clustered in terms of their patterns of care to highlight potential relationships between those patterns and clinically significant outcomes, such as event-free survival. These types of analyses are aimed at properly stratifying a group of patients, verifying guideline adherence and their impact, and giving better structure to real-world outcomes. For example, Baker et al. propose an approach to extract clinical pathways, or *careflows*, of breast cancer patients during chemotherapy by using data routinely collected in the electronic health record (EHR) [3]. The proposed methodology, based on Markov models, is able to highlight the complexity of real-life pathways with respect to the ideal ones proposed in clinical guidelines, identifying a set of critical situations, such as readmission to the hospital. This approach can help to identify unmet needs that can cause non-compliances to the guidelines. An interesting and broadly applicable approach to examining careflows in detail, reflecting the temporal nature of the clinical events that compose them, is to use the *CareFlow Mining* (*CFM*) approach that was recently developed by the authors of this paper, and successfully applied to different clinical settings [4,5]. CFM can be used to extract emerging temporal patterns of clinical diagnoses and procedures from long sequences of events, which cannot be retrieved by resorting to traditional SQL queries. CFM results in time-oriented patients' stratification, which might be related to significant clinical outcomes, and thus used to define *temporal phenotypes*.

An important source of information for CFM is represented by administrative data. In this work, we extend the work already presented in [5] to explore the potential of using administrative data in gaining

* Corresponding author.
E-mail address: lucia.sacchi@unipv.it (L. Sacchi).

https://doi.org/10.1016/j.artmed.2020.101855
Received 6 September 2019; Received in revised form 1 April 2020; Accepted 2 April 2020

insight into different care scenarios that occur in a hospital breast cancer unit. Administrative data are collected and exploited for billing purposes, but they also reflect the medical actions performed to address specific health conditions. Those data have the advantage of being structured, time-stamped, and less prone to missingness than clinical data. On the other hand, they are less informative, as they just carry the information that a specific action has been performed, without reporting a clinical observation or outcome. Administrative data are also very granular, thus implying that data related to similar clinical scenarios may be represented by different codes. If not properly managed, this variability can cause CFM algorithms to extract models [6] that are not usefully interpretable [6].

To guide the creation of a more comprehensible process model, explicit definitions of key clinical activities is crucial. This can be done following two strategies: on the one hand, it is possible to manually classify the events relying on a knowledge-driven approach supported by domain experts [7]. On the other hand, it is possible to use automated algorithms aimed at grouping similar clinical events into relevant categories [8–10]. When clinical events are expressed in natural language, *Topic Modeling* (*TM*) techniques can be applied as a preprocessing step in process mining [9–11]. TM is a text mining methodology to cluster documents on the basis of their content. TM algorithms, such as Latent Dirichlet Allocation (LDA) [12], use a probabilistic approach to discover themes (or topics) in large archives of documents and automatically annotate them, without any need of prior labeling.

As mentioned, some works in the literature have already exploited TM as a first step for clinical pathway mining [9,10]. In [9], the authors use TM to synthesize the daily activities of patients, incorporating the clinical events into a sequence of topics computed for each day. They then apply a process mining algorithm to demonstrate that the created sequences of topics are clinically meaningful. The focus of this paper is on the optimization of LDA when applied to clinical data, and the final process mining step serves to assess the quality of the proposed topics, rather than stratifying the population into clinically relevant subgroups. The same approach can be found in [10], where the authors present further experiments on TM, focusing on expanding the functionalities of LDA by embedding constraints in the construction of the stochastic model, to enforce the meaningfulness of the discovered topics.

In this work, we propose an analytic pipeline based on a combination of TM and CFM, which will advance previous research in several directions. First, we have focused on improving our CFM algorithm to be able to consider simultaneous events carried out during the same hospitalization. Furthermore, we have performed an evaluation of the clinical relevance of the results by comparing the extracted careflows in terms of clinical outcome, which was not available in [5]. Finally, with respect to the other approaches presented in the literature, this work has the main goal of showing a complete analysis workflow that, starting from administrative and clinical data of breast cancer patients, has the main goal of extracting clinically relevant temporal phenotypes to stratify patients according to their flows of care.

We show results on a dataset of more than 3000 patients who underwent breast surgery at the hospital IRCCS ICS Maugeri of Pavia (ICSM), Italy. Two data sources were used: Hospital Information Systems (HIS) recorded procedures, used for billing and administrative purposes, and a registry of clinical and molecular data collected by the Oncology Ward service. The events used for CFM are derived from ICD9-CM procedures, which is the standard coding system used by administrative information systems in Italian hospitals. Through rigorous examination of the data, we describe the disease evolution patterns of breast cancer patients treated at ICSM. The mined patterns show clinical significance in terms of specific clinical endpoints, such as recurrence or metastases.

2. Methods

According to clinical guidelines [2], the treatment of a breast cancer patient after the first surgery proceeds through a series of hospitalizations and Short Procedure Unit (SPU) visits, which are aimed at delivering the therapy, performing additional surgical interventions (e.g., reconstruction, treatment of relapse, etc.) and examinations, or dealing with possible complications of the disease or treatment. Each hospitalization is in turn characterized by a variable number of procedures, which are carried out on a patient in the period that goes from admission to discharge. This creates two temporal dimensions, the first one that represents the overall flow of the hospitalizations, and the other that represents the inner flow of procedures within a single hospitalization. From the perspective of understanding the main careflows enacted in a specific institution, the temporal trajectory that needs to be considered is the one related to the sequence of hospitalizations. Therefore, it is more important to preserve and synthesize the information on the clinical procedures performed during hospitalization-specific events, rather than their temporal occurrence within the hospitalization.

To tackle this twofold problem, we propose a pipeline that is based on topic modelling and careflow mining. Specifically, the main care processes are discovered by performing CFM on events extracted using topic modelling to summarize the information related to within-hospitalization procedures. Fig. 1 explains in more detail the steps that are carried out:

1 Starting from a cohort of cancer patients who underwent breast surgery, we extract the procedures codes related to all the hospitalizations following the first surgery.
2 To summarize the groups of procedures included in each hospitalization, we apply the TM step. In this step, we process the description of the codes related to the set of procedures included in a single hospitalization record as a document and assign a topic to it. We then create the event log, where each event is the topic assigned to the considered hospitalization, and the timestamp of the event corresponds to the time span of the hospitalization.
3 The CFM algorithm [5] is run on the topic-based event log derived as described in Step 2. With the help of expert physicians, the careflows are further summarized into clinically meaningful temporal phenotypes.
4 Temporal phenotypes are compared in terms of clinical endpoints and used in multivariate models including relevant clinical information about the oncologic disease.

2.1. Study cohort

Data were retrospectively collected on patients treated at the Breast Unit of ICSM, a high-volume tertiary centre directly involved in extensive mammographic screening programs in northern Italy. We initially considered patients included in a manually curated dataset maintained at the Breast Unit, which includes data on 3564 subjects followed from January 2007 to November 2018 [13]. The inclusion criteria were: i) a confirmed diagnosis of breast cancer and ii) one or more following surgery procedures of any kind (e.g., lumpectomy, mastectomy, nipple-sparing mastectomy, skin-sparing mastectomy with reconstruction). Exclusion criteria were: i) distant metastases at diagnosis, ii) a previous diagnosis of cancer (including breast cancer), iii) benign breast diseases.

To build the patients' sequences of events, we extracted from ICSM HIS all the inpatient hospitalizations and SPU visits that were performed after the first diagnosis of breast cancer. Since we were interested in activities performed during inpatient hospitalizations, we extracted the billing data related to the ICD9-CM codes of the procedures registered in the discharge summary. For SPU visits this information is not available, as these visits are not coupled with a discharge summary

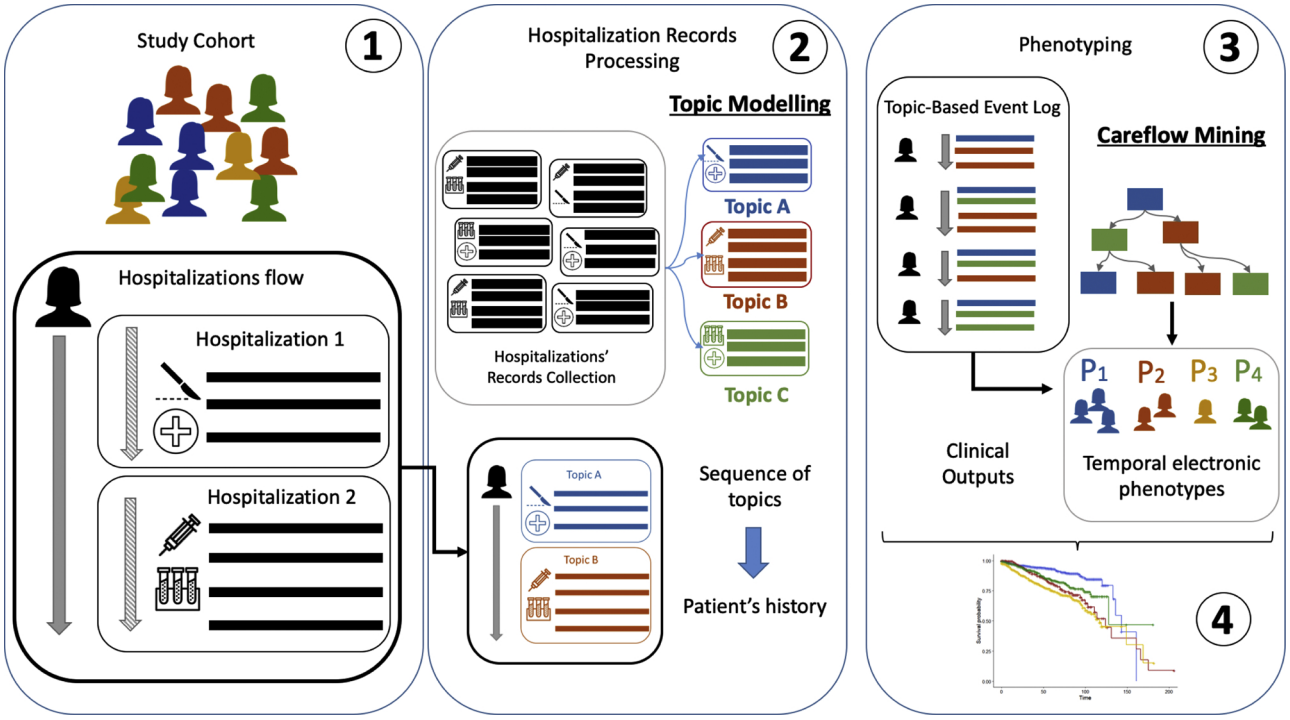


Fig. 1. The proposed analytic pipeline: 1. Starting from a cohort of cancer patients who underwent breast surgery, we extract the procedure codes related to all the hospitalizations following the first surgery. The sequences of procedures allow identification of two temporal dimensions: (i) the overall hospitalizations flow (i.e. from the first surgery to the last recorded hospitalization and clinical outcome), and (ii) the inner flow of procedures within the single hospitalizations. 2.To summarize the groups of procedures included in each single hospitalization record, we apply the TM step. In this step, we process the description of the codes related to the set of procedures included in each hospitalization record as a document and assign a topic to it. We then create the event log, where each event is the topic assigned to the considered hospitalization, and the timestamp of the event corresponds to the time span of the hospitalization. 3. The CFM algorithm is run on the topic-based event log derived as described in Step 2. The careflows are further summarized into clinically meaningful temporal phenotypes (P1,..P4) with the help of medical experts. 4. Temporal phenotypes are compared in terms of clinical endpoints and used in multivariate models.

but rather with a textual medical report. For this reason, these visits are simply recorded with their occurrence date, without any other information. Procedures carried out during SPU visits are usually follow-up encounters or administration of treatments such as radiotherapy and hormone therapy.

To extract data from the HIS, we used Pentaho Kettle [14], a Java-based open source platform for extract-transform-load (ETL) procedures. Data were exported in. csv format, and the following analysis steps were carried out using R [15]. From the original set of patients, we excluded 218 subjects who did not undergo any procedure. The following analysis was then performed on data from 3346 patients.

The manually curated dataset included baseline patients' characteristics and survival endpoints at the last available follow-up. We took into consideration Age at diagnosis (years), Type of surgery (Lumpectomy, Mastectomy), Histological Type (DCIS, Ductal invasive carcinoma, Lobular invasive carcinoma), Grading (G1, G2, G3) Staging (0-III), Biomolecular subtype (Luminal A, Luminal B, HER2+, TNBC), Hormone therapy, Chemotherapy, Radiation therapy, Neoadjuvant therapy and Lipofilling intervention. The clinical endpoints taken into consideration were: 1) the 10-year loco-regional recurrence (LRR)-free survival probability and 2) the 10-year distant metastases (DM)-free survival probability 3) the 10-year overall survival probability.

2.2. Topic modelling

The aim of the TM step is to represent each hospitalization using single labels that synthesize the information related to the procedures performed on the patient. To perform this step, we exploited the Latent Dirichlet Allocation (LDA) method [16], which is a widely used method for TM, and is implemented in the R package 'topicmodels' [17].

LDA is an unsupervised algorithm that allows clustering a set of documents (document corpus) into K different topics, where K is a user-defined parameter that fixes the set of topics, each of which represents a set of words. The goal of LDA is to map all the documents to the topics, such that the words in each document are mostly captured by the established K topics.

The LDA algorithm is a generative probabilistic model where:

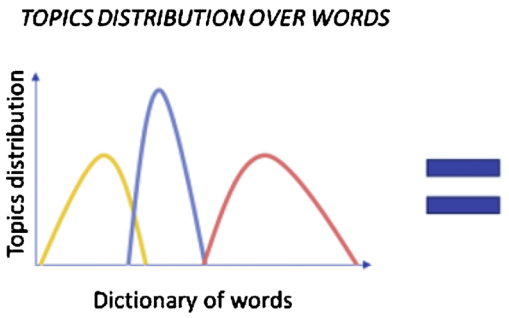
(i) Each latent topic can be described by a probability distribution over a dictionary of words. This dictionary is composed by all the words in the documents included in the corpus. The distribution of topic over words is indicated by the matrix ϕ, which is graphically illustrated in Fig. 2, where the columns correspond to topics, and the rows to words. Each column of ϕ represents the probability distribution of the words, given the topic. On its turn, each column of ϕ is a random variable with a Dirichlet distribution.

(ii) Each document is represented by a random mixture over topics. Such mixture is described by a latent variable vector z of dimension K; the probability distribution of z is multinomial with parameter θ. θ is a random variable with Dirichlet probability distribution.

Given a corpus D of documents, each document d having N_d words, generated by a single topic from a set of K topics. Bayesian estimation allows the derivation of the posterior probability of the latent variables z, θ and φ. The posterior moments can be used to derive the point estimates for all the latent variables.

To apply LDA, we created a document corpus where each document refers to a single hospitalization, and the content of the document is the list of procedures carried out during that hospitalization. We pre-processed the corpus to remove punctuation, numbers, and stop words.

Modifying the LDA algorithm embedding internal constraints was not the goal of this study, but, taking the example from the study in



Dictionary of Words	Topic 1	Topic 2	Topic 3
w^1	0	0.1	0
w^2	0	0.2	0
w^3	0	0.5	0
w^4	0	0.2	0.3
w^5	0	0	0.7
w^6	0.1	0	0
w^7	0.3	0	0
w^8	0.4	0	0
w^9	0.2	0	0

Fig. 2. Graphical visualization of the topic distribution over words, and the ϕ matrix showing the values of words probabilities for each topic.

[10], we defined a method to ensure that the model we selected is coherent and develops a meaningful set of topics. In particular, to evaluate the quality of the models generated via LDA and to select the optimal number of topics K, we introduce the following scores: *consistency*, *redundancy*, *importance,* and *perplexity*.

Consistency, redundancy, and importance are heuristic parameters, which we defined as reported in the following of this paragraph. The perplexity measure is a standard value used to evaluate the LDA result [12], and it is in general related to the robustness of topic assignation.

Consistency and redundancy are related to the words defining a topic and are computed on the basis of words repetitions in a topic. Consistency expresses the reproducibility of topics composition after multiple runs of the same model with the same K. Redundancy indicates how many words characterizing a topic are also present in other topics of the same model. To define these two indicators, we introduced the concept of *"topic words"* as the smallest amount of words (in descending order of importance) that are needed to reach the 80 % of the cumulative probability of φ_k, i.e. the words probability distribution of the k-th topic. Fig. 3 reports an example of this definition, where the set of topic words for Topic 1 is represented by words w^1, w^2 and w^3, whereas the topic words for topic 2 are w^5 and w^6.

Consistency

Since model creation is a probabilistic method, there is the possibility that each run of the algorithm can generate a different configuration of topics. To overcome this issue, we decided to run the algorithm multiple times and check the consistency C of each generated topic T, counting if and how T and its words will be present in another iteration of the model.

$$C(T) = \sum_{i}^{I\_max} (\max_{1\le k\le K} (\bigcap_{words} (T, t_{i,k})))/I\_max$$

Where I_max is the total number of iterations of the LDA model, K is the number of topics, $t_{i,k}$ is the *k-th* topic of the *i-th* run and $\bigcap_{words}(t1, t2)$ is the number of *"topic words"* shared by t1 and t2. The overall *Consistency* of a model is calculated as the mean of the consistency of all its topics.

Redundancy

Another problem that could arise by running LDA is that each word can potentially have non-zero probability to generate more than one topic. For this reason, for each topic T in a model, we compute the redundancy R by evaluating how many words are repeated in other topics of the model. R is computed as follows:

$$R(T) = \max_{\substack{1\le k\le K \\ T\neq t_k}} (\bigcap_{words} (T, t_k))/length(T)$$

Where is the total number of topics, t_k is the *k-th* topic of the LDA model (note that topic T will not be compared with itself), $\bigcap_{words}(T, t_k)$ is the

TOPICS	Topic Words	Words probability	# of "important words"
Topic 1	w^1	**0.4**	3
	w^2	**0.3**	
	w^3	**0.2**	
	w^4	0.08	
	...	...	
Topic 2	w^5	**0.7**	2
	w^6	**0.1**	
	w^7	0.1	
	...	...	
Other topics	...	...	

Fig. 3. Definition of the topic words.

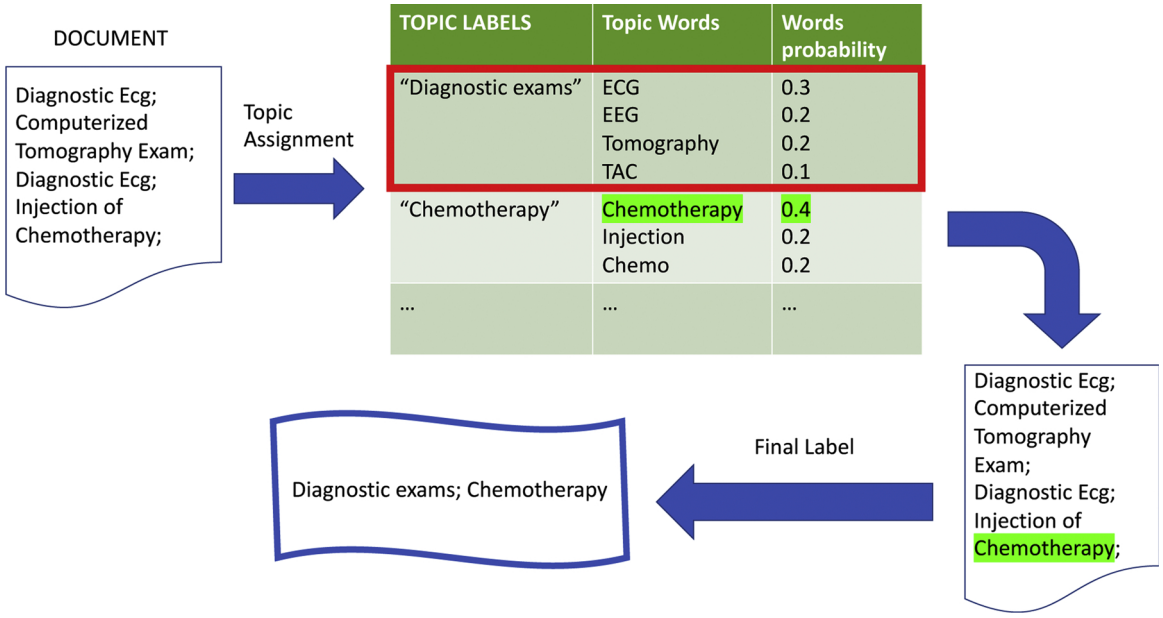


Fig. 4. The document meets the labeled topic model and it is assigned to "Diagnostic Exam", because of the prevalence in the text of words belonging to this topic. Alongside the topic label, the final document label is enriched with another word.

number of *"topic words"* shared by T and t_k, and $length(T)$ is the number of *"topic words"* for topic T. The *Redundancy* of the overall model is calculated as the mean of the redundancy of its K topics.

Importance

The importance of a topic for a document measures if the words belonging to that document are mostly extracted from that topic.

The overall importance I of a topic T can be thus calculated as follows:

$$I(T) = \left(\sum_{d}^{D_T} \theta_T(d) - \left(\max_{\substack{1 \le k \le K \\ k \ne T}} (\theta_k(d)) \right) \right) / length(D_T)$$

Where D_T is the group of documents assigned to topic T, t_k is the *k-th* topic of the LDA model (note that the topic T will not be compared to itself), $\theta_k(d)$ is the probability of the k-th topic for the document d, and $length(D_T)$ is the number of documents assigned to topic T. The *Importance* of the topic model can be computed as the mean of the importance of its topics.

Perplexity

Perplexity is a commonly used score to evaluate the LDA models on held-out data. It was also used in clinical applications to select the appropriate number of topics (K) [10] to identify concise and interpretable process models. Perplexity is inversely correlated to the generalizability of the model, as it evaluates the likelihood of the proposed assignation of topics to words and documents. A trade-off between K and perplexity has to be found in order to respond to the required generalization capabilities of fitted models. The perplexity score can be computed with the built-in function 'perplexity' of the R package 'topicmodels' [17]. Formally, for a corpus D, the perplexity is defined as follows [12]:

$$Perplexity(D) = exp\left\{ -\frac{\sum_{d=1}^{D} \log(p(d))}{\sum_{d=1}^{D} N_d} \right\}$$

Where log(p(d)) is the log likelihood of the model for a document d, D is the corpus of documents and N_d is the number of words of the document d.

Selection of K, number of topics

To select the best model, we ran the algorithm multiple times with different values of K, and evaluated both the defined scores and the content of the discovered topics. We selected the model that allowed obtaining a good trade-off between high values of Consistency and Importance and low values of Redundancy and Perplexity.

Topic labelling

Once the LDA topic model is selected, it provides a list of K topics, each of which is characterized by a list of *"topic words"*. Since the model doesn't provide labels to synthesize each topic, as they depend by the informative content of the topic itself, a manual labelling step is needed. This is a procedure that is frequently performed to be able to use the topic modeling results for further analysis steps [18]. In this paper, we considered the list of *"topic words"* for each topic, and defined a meaningful name for the topic.

Document labelling

Once the process of topic model selection and labeling is complete, the next step is to tag each document (i.e. each hospitalization) with the label of the topic to which the document has been assigned with the highest probability.

In some cases, the assignment of a document to a topic might only partially represent the information content of the document itself, due to the high number of procedures performed during the same hospitalization. In order to tackle this issue, we represent each document both with the label of the topic it has been assigned to and with the word in the document with the highest probability in matrix ϕ, considering the entire topic set. This document-specific word can confirm or enrich the meaning related to the topic label.

If we consider for example the document shown in Fig. 4, we can see that it is assigned to the topic labeled as *Diagnostic Exams*. Considering the words in the document and the probabilities related to them, though, we can see that the word with the highest score is Chemotherapy. Even if this word is not specific to the topic *Diagnostic Exams*, we include it in the label of the document, to enhance the representativeness of the label. As we will show, the most important word in the topic is most frequently in accordance with the overall label of the topic. In other cases, using the first word is a good strategy to improve the interpretability of the document.

2.3. Careflow mining

We applied the algorithm described in [4] and implemented in [19]. The algorithm considers the temporal nature of the data, mining the most frequent careflows in terms of process events, which in this work are represented by hospitalizations and SPUs summarized by the Topics identified in the previous step.

The CFM algorithm works on a file with a list of ordered events, where each row includes the following information:

- ID: the patient subject to the event
- EVENT: name of the event, in our case the label assigned to each hospitalization/SPU visit at the end of the TM step
- DATE_INI: event start date - the hospitalization admission
- DATE_END: event end date - the discharge

The algorithm extracts frequent careflows from process data and it is inspired by sequential pattern mining techniques. It discovers frequent careflows, where frequency is defined in terms of *support*. The support (S) of a careflow is defined as the number of patients (Ns) who undergo the sequence of events by which it is composed divided by the total number of patients in the analysed population (N).

$$support\,(S) = \frac{Ns}{N}$$

Frequent sequences are those that have a support S greater or equal than a user-defined threshold. Thresholds are used to guide the search process such that only the most frequent patterns are extracted. The algorithm works starting on the first events of the patients' sequences and selecting those that are more frequent than a pre-defined threshold on support (min_support). The algorithm adds steps to the careflows by iterating the support computation on the events that follow the initial set, until no more frequent sequences can be extracted, or a maximum number of events is reached. This second constraint can be controlled by another parameter called max_length. This discovery step of the algorithm requires a careful assessment regarding min_support and max_length. The effect of these two parameters affects the generalization and precision of the CFM models: low min_support and high max_length might lead to overfitting and a difficult interpretation of the results, losing power to summarize patients' care pathways. On the other hand, high min_support and low max_length can retain only a general description of the initial events of the majority of patients, loosing details and becoming under-fitted.

To select the min_support and max_length parameters, we followed the strategy described in [4] and we performed a grid search by varying these parameters in a defined range. At each iteration, we computed the number of extracted careflows, the average number of patients per careflow, the average number of events not represented in the final careflows (missed events), and the proportion of patient sequences fully represented by the mined careflows (true match rate). We decided to use the parameters that resulted in the best trade-off among the considered indicators. Maximizing the true match rate allows maximizing the homogeneity of the sequences that are included in the same careflow.

The result of the algorithm can be represented using a Directed Acyclic Graph (DAG), where nodes are the events, while arcs represent temporal connections among them. The resulting DAG is enriched by temporal information. In particular, for each event of a careflow, the graph shows the number of patients undergoing the event, and the median, 25^{th}, and 75^{th} percentile of the duration of that events for the patients who verify it. The arcs report the same statistics, computed on the duration of the transition between the two events that are connected by the arc. In the final event for each careflow, the total history time for the patients of the careflow is provided, as median of times between first and last displayed events of the careflow.

2.3.1. Careflow assessment

The careflows resulting from the application of the CFM algorithm are able to divide and separate the population of patients into a set of sub-cohorts, which can help identify different temporal phenotypes.

In order to make careflows comparable from a statistical point of view and derive phenotypes that have a clinical meaning, CFM results have been presented to expert clinicians who – thanks to their expertise and knowledge on the real processes - validated them into specific phenotypes. As a result of this process, some of the careflows originally resulting from the application of the CFM algorithm were merged. It has to be noted that this re-grouping was completely guided by the CFM results: clinicians suggested group-specific branches of the DAG given their knowledge that the procedures mined in the processes were equivalent or that were the exact same procedure recorded in different ways. The so-derived phenotypes can be further analysed and assessed in terms of clinical outcome and disease evolution.

In this work, the extracted temporal phenotypes have been characterized in terms of the following clinical endpoints: Local/loco-regional recurrence (LLR), Distant recurrence Metastasis (DM), and Death. Survival probabilities are estimated by Kaplan-Meier methods. Groups were compared by multivariate Cox proportional hazards regression model with stepwise model selection by AIC, including variables associated with the outcomes, such as the type of therapy (hormone therapy, chemotherapy and radiation therapy) and biomolecular subtype to avoid biases. A multivariate survival analysis was performed excluding those subjects for which the right censoring follow-up date preceded the last mined event in the CFM. Statistical significance was set at $p < 0.05$ (two-tailed). Data analysis was performed using the R software.

As a final step, to evaluate the described pipeline with respect to our previous work, we have compared the results obtained with the approach proposed in this paper with those obtained by applying to our data the approach originally presented in [5]. This previous approach applies the CFM algorithm to data where single events are characterized on the basis of the type of hospitalization (Admission or Short Procedure Unit) and the ward, without applying the topic modelling step.

3. Results

3.1. Data pre-processing and preparation

Table 1 summarizes the demographic and clinical characteristics of the study cohort. The statistics of the population in terms of hospitalizations, SPU visits and procedures are shown in Table 2.

Clinical and administrative data were integrated using unique patient identifiers. Left-censoring was performed considering the first breast surgery (also reported as baseline time point in the clinical data stream) and its related procedures (e.g. biopsies or neoadjuvant therapies administrated before the surgery) registered in the HIS. Right-censoring was performed considering for each subject the last procedure registered in the HIS at the moment of data extraction, and compared with the clinical endpoints, in order to exclude subjects whose careflow events overlap or follow the clinical outcome.

In addition to using the low-granularity set of the 363 ICD9-CM codes available in the data, we also mapped the codes into higher-level categories. Pre-processing techniques are available for this purpose, like the Clinical Classification Software (CCS) which provides a suitable categorization scheme for ICD procedures [20]. Since the CCS is structured on several levels of granularity, we started with a mapping based on a finer CCS level (level 2) for breast-related procedures, and a coarser level (level 1) for non-breast-related procedures. After a further step of manual review, we obtained 36 Procedures. The ICD9-9 CM codes/Mapped Procedures mapping is provided in the Supplementary section (Appendix 1).

Clinical events were thus represented as: (i) sequences of ICD9-CM

Table 1
Baseline clinical and epidemiological data.

TOTAL	3346
Age Mean (SD)	58.74 (13.38)
Type of Surgery N (%)	
Lumpectomy	2328 (70.31)
Mastectomy	1018 (30.75)
Histological Type N (%)	
DCIS In situ cancer	337 (10.18)
CDI Invasive ductal cancer	2457 (74.21)
CLI Invasive lobular cancer	552 (16.67)
Staging N (%)	
Stage 0	439 (13.26)
Stage 1	1670 (50.44)
Stage 2	908 (27.42)
Stage 3	329 (9.94)
Grading N (%)	
G1	427 (12.9)
G2	1934 (58.41)
G3	985 (29.75)
Hormone therapy N (%)	
No	767 (23.17)
Yes	2579 (77.89)
Chemotherapy N (%)	
No	2084 (62.94)
Yes	1262 (38.12)
Radiotherapy N (%)	
No	1088 (32.86)
Yes	2258 (68.2)
Neo adjuvant chemotherapy N (%)	
No	2972 (89.76)
Yes	337 (10.18)
Unknown	37 (1.12)
Cancer multifocality N (%)	
No	2528 (76.35)
Yes	818 (24.71)
Bio molecular subtype N (%)	
Luminal A	1823 (55.06)
Luminal B	1044 (31.53)
Her2+	175 (5.29)
TNBC	304 (9.18)
Lipofilling intervention N (%)	
No	2907 (87.8)
Yes	439 (13.26)
Local/loco-regional recurrence (Yes) N (%)	195 (5.89)
Metastasis - Distant recurrence (Yes) N (%)	260 (7.85)
Death (Yes) N (%)	176 (5.32)

Table 2
Hospitalizations, SPU visits, and procedures.

Total number of hospitalizations	8387
Total number of procedures	20765
Total number of SPU visits	1650
Distinct procedures (ICD9-CM codes)	363
Average number of hospitalizations per patient (SD)	2.55 (2.47)
Average number of SPU visits per patient (SD)	2.13 (1.38)
Average number of procedures per patient (SD)	6.33(7.10)
Average number of distinct procedures per patient (SD)	4.90(4.00)
Average number of procedures per hospitalization (SD)	2.47 (1.45)
Average number of unique procedures per hospitalization (SD)	2.33 (1.31)

procedures' codes (2011 version) within each hospitalization, and (ii) sequences of mapped procedures derived from the CCS and the oncologist manual revision. These events were used as inputs to the Topic Modeling step and the results of the two approaches were compared, as described in the following section.

3.2. Topic modeling

On the basis of the two clinical events representation, the analyses were performed on two different documents corpora: (i) in the first document corpus each document is the set of the Italian description of ICD9-CM procedures performed in one hospitalization, (ii) in the second document corpus each document is built in the same way, but we used the CCS and manual oncologist reclassification in the English language. Due to the difference of the vocabulary and the language used in each corpus, we performed two different preprocessing and cleaning steps. We report in the following the detailed results obtained using ICD9-CM codes, whereas for the CSS based approach we herein show only a summary of the main findings (the detailed results are reported in Appendix 4).

3.2.1. TM of ICD9-CM procedures

After the creation of the corpus, stop words, punctuation and numbers were excluded. We used the default Italian stop words list provided by R package 'topicmodels' version 0.2–8 (see Supplementary material – Appendix 2). An additional set of stop words was manually added to complete the list. This set included words that represent generic terms in the Italian language ("altro", i.e.,"Other"), specific terms that are not important in our case study ("arterioso", "microscopico", i.e, "arterial","microscopic"), specific terms related to our case study but too frequent ("mammella", i.e.,"breast") or applied to different concepts ("iniezione", "infusione", i.e.,"injection" and "infusion", words that are present in the description of both chemotherapy and other therapeutic administration). Since the list of stop words could cause an entire document to be deleted, the stop words must be carefully chosen in order to avoid it.

After the cleaning step and the removal of the stop words, the dictionary reduced in size from 588 to 498 words. On average, a single word is present in 91 documents (sd = 317). The most frequent word ("Asportazione", i.e., "Removal") is present in 3061 documents. A document is composed on average of 6 words (sd = 3.1), from a minimum of one to a maximum of 40 words. The 75 % of the documents contain 8 words or less.

As reported in the Methods section, to choose the topic number we ran the LDA algorithm considering different values for k, ranging from a minimum of 2 topics to a maximum of 16 topics.

As already explained, in order to choose k, we took into consideration *consistency*, *redundancy*, *importance,* and *perplexity* (see Fig. 5).

Consistency does not significantly vary with K and it is always greater than 0.5. According to the definition reported in Section 2.2, this means that, on average, half of the words representative for a topic are preserved over repeated LDA runs. After an initial decrease, Redundancy increases for $K >= 4$, reaching a plateau around 0.25. This value means that, on average, only the 25 % of the topics are overlapping in a single LDA run. Importance increases for $K = 3$, and it is always higher than 0.5, meaning that, on average, the difference between the probability of the most important topic and the second most important topic for a document is 0.5. Finally, perplexity decreases until $K = 5$, and then remains almost stable. This behavior was expected, as the initial increase of the number of topics helps the model to provide stronger assignation of documents to the topics, thus causing perplexity to rapidly decrease. When K gets higher, the benefit of increasing the number of topics becomes less effective, as it leads to lower probabilities of assignation.

Looking at these indexes and at the topics generated from the different set-ups, we selected $K = 6$, as a compromise between the capability of describing the domain complexity and the need to group procedures in a clinically coherent way. The extracted topics, the topic words set, and the topic labels that were assigned to each group are shown in Table 3.

Topic 1 includes words related to lymph nodes operations and/or skin graft procedures. Topic 2 represents plastic surgery and it includes the placement or removal of prosthesis, either unilateral or bilateral, plastic surgery mammoplasty associated with mastopexy and other

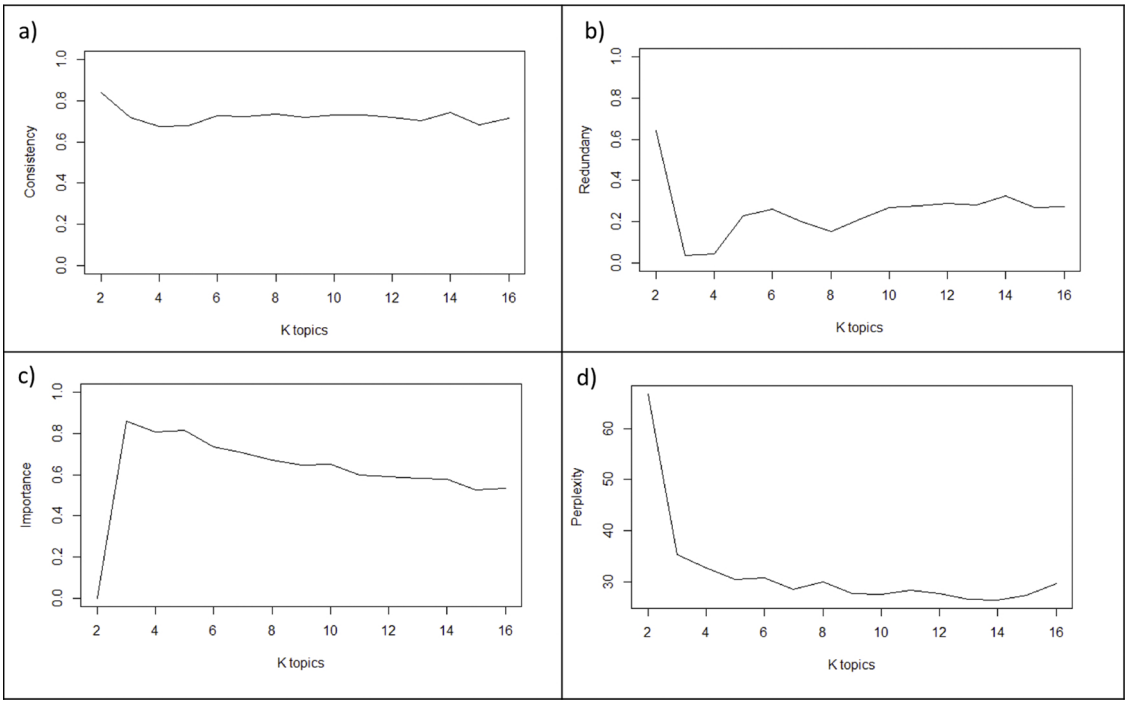


Fig. 5. Consistency (a), Redundancy (b), Importance (c) and Perplexity (d) values on varying k, when TM is performed on ICD9-CM codes.

corrective surgery techniques. Topic 3 is the most heterogeneous one, as shown by the variety of the exams and procedures included in its word set, which is also characterized by a large number of words (45). The topic can be interpreted as the set of the exams performed on a patient during the oncology process of care, starting from diagnostic exams and imaging localization of the tumor and ending with assessment and rehabilitation exercise performed during follow-up. Topic 4 represents chemotherapy and other aspects related to it. Of note, this topic also includes two words related to ultrasound examinations of the heart and electrocardiogram: chemotherapy is often related to cardiological tests such as echocardiogram, because chemotherapy drugs could be cardiotoxic and so patients should undergo cardiological monitoring.

Topics 5 and 6 represent surgical interventions on the breast. Topic 5 includes words related to breast-conserving surgery. In particular, the most representative words describe the surgical procedure, which is usually composed of tumor localization through radioisotopes, quadrantectomy and lymph nodes removal. Topic 6 is related to mastectomy, which is usually coupled to the positioning of a breast insert to prepare the patient for future reconstruction. Underarm lymph nodes may or may not be removed depending on the situation.

3.2.2. TM of remapped procedures

To apply TM on the procedures remapped according to the strategy explained in Section 3.1 (based on CCS and manual oncologist reclassification), we used a similar strategy as the one described for raw ICD9-CM. Topics, their topic word sets, and the Topic labels are shown in Appendix 4. Even if the two document corpora are different, the results found on the TM scores for the remapped procedures indicate that it is possible to choose the same topic number, $K = 6$. Moreover, the clinical interpretation of the topics is almost the same as the one presented in Table 3. As a consequence, we may argue that, in this study, the TM approach is robust in highlighting the most relevant clinical conditions, independently from the language and the procedure representation system used.

Given the substantial overlap of the results of the two TM strategies, in the following we will report the CFM results obtained after *TM* of the *ICD9-CM procedures* step. This approach relies on standard coding and requires less pre-processing than manual remapping, thus being more generalizable to other clinical contexts.

3.3. Careflow mining

The CFM algorithm has been applied to the event logs derived from time-ordered sequence of hospitalizations of each patient. CFM algorithm parameters were selected following the grid-search approach presented in the Methods section. In particular, we performed a grid search by varying min_support in the range 2–50 and max_length in the range 3-10. Fig. 6 shows the number of careflows (blue), the average number of patients per careflow (purple), the average number of missed events (green), and the true match rate (red) for each value of the pair of parameters (numeric values are reported in Appendix 6). As shown in Fig. 6, the parameter values max_length = 10 and min_support = 10 result in a relatively low number of detected careflows, while preserving a good matching rate and a low number of missed events. We used these values in the following analyses.

The application of the CFM algorithm with the selected parameters resulted in a total of 160 events that were organized in 81 careflows. Table 4 reports the list of the 19 distinct events included in the careflows. As explained in the Methods section, the events' labels contain the topic name and the most informative word of the document. The longest careflows comprise of 5 events, and an average history length of 3.33 (SD = 0.9) and a median equal to 3. The complete list of the extracted careflows is reported in the Supplementary (Appendix 5).

Considering the resulting careflows, we have been able to further reduce the number of sub-groups by merging sequences of events with the same meaning. For example, sequences including multiple occurrences of plastic surgeries were grouped into a single group (Cluster 1, plastic surgery). From the initial 81 histories we have derived 9 clusters, as follows:

- Cluster 1: Reconstruction/plastic surgery. This cluster contains histories related to cases characterized by one or more occurrences of plastic surgery for breast reconstruction. These patients are referred to ICSM after a first intervention that was carried out in another hospital.

Table 3
Topic selected with k = 6 when using ICD9-CM procedures. The table reports the words included in the topic, the number of important words (i.e. words to reach a total probability > 0.8) and the assigned topic label.

	ICD9-CM procedures		
TOPIC	ITALIAN WORDS (TRANSLATION)	# Important Words	TOPIC LABEL
1	Cute (Skin) Sedi (Sites) Linfonodi (Lymph nodes) Innesto (Graft) Ascellari (Axillary) Radicale (Radical)	6	Skin graft / Lymph nodes operations
2	Protesi (Prosthesis) Impianto (Implant) Monolaterale (Monolateral) Bilaterale (Bilateral) Mastopessi (Mastopexy) Riduttiva (Reductive) Mammoplastica (Mammoplasty) Rimozione (Removal)	8	Plastic
3	Ecografica (Sonographic) Torace (Thorax) Terapeutiche (Therapeutic) Tomografia (Tomography) Esercizi (Exercise) Radiografia (X-Ray) TAC (CT) Valutazione (Evaluation) Esami (Exams) Elettrocardiogramma (ECG) …	45	Other exams and therapies
4	Tumore (Tumor) Chemioterapeutiche (Chemotherapeutic) Esami (Exams) Antineoplastico (Antineoplastic) Iniezione infusione (Injection Infusion) Ecografica (Sonographic) Elettrocardiogramma (ECG)	7	Chemotherapy
5	Asportazione (Removal) Quadrantectomia (Quadrantectomy) Scintigrafia (Scintigraphy) Linfatico (Lymphatic - singular) Linfatiche (Lympahtic - plural) Radioisotopi (Radioisotope)	6	Lumpectomy
6	Monolaterale (Monolateral) Inserzione (Insertion) Mastectomia (Mastectomy) Espansore (Expander) Tissutale (Tissue) Asportazione (Removal) Linfatico (Lymphatic) Scintigrafia (Scintigraphy) Mammectomia (Mammectomy)	9	Mastectomy

- Cluster 2: Surgery + Therapy. The cluster maps well to one of the breast cancer guidelines of surgery and therapy.
- Cluster 3: Surgery + Therapy + Plastic Surgery. The third cluster adds plastic surgery to the previous careflow cluster.
- Cluster 4: double surgery. This group includes all double surgery cases. It has been further split into two sub-clusters on the basis of the time span between the two surgeries. In cluster 4a, the second surgery occurred very close (within two months) after the first one: this is likely related to a second intervention decided after having considered the histopathological exams of the breast samples collected by a first conservative intervention. Cluster 4b reports cases related to a second surgery that was performed later in time. This may correspond to a second cancer episode or to a recurrence of the same one, such as local or regional recurrences.
- Cluster 5: Surgery + Plastic Surgery. Patients in this cluster are managed by ICSM only for their surgical intervention, whereas therapy is decided and administered in another hospital.
- Cluster 6: Neoadjuvant therapy. This cluster includes patients for which a SPU visit is performed before the first surgery. This is related to patients who undergo neoadjuvant therapy, which consists in the administration of therapeutic agents, such as chemotherapy or hormonal therapy, before the main treatment. This is confirmed by the clinical data that are associated to this group: 96 % of the patients underwent such treatment.
- Cluster 7: surgery + rehabilitation. Patients belonging to this group underwent surgery followed by rehabilitation at the hospital.
- Cluster 8: surgery. Patients in this group had only undergone surgery at ICSM. This is largest group, including nearly half of the studied population.
- Cluster 9: surgery + exams. Patients belonging to this group underwent surgery (as only form of treatment) and further exams to investigate the clinical outcome of surgery.

Fig. 7 illustrates the original CFM results and the regrouping, performed by expert clinicians, into Clusters 2 (Surgery and therapy), 4 (double surgery) and 8 (surgery). Events labelled as “Day Hospital; SPU visit” in Cluster 2 indicate one or more chemotherapy treatments.

It’s interesting to note that, since 2012, the administrative management of chemotherapies has changed. While before 2012 chemotherapies were performed during one-day hospitalizations characterized by a regular discharge letter with the indication of ICD9-CM procedures, after 2012 chemotherapy has been managed as an outpatient service, which has just a textual report as a result. The information related to that kind of visit is preserved, as we merged careflows assigned to the topic chemotherapy to careflows assigned to the topic day-hospital.

Table 5 reports the clusters’ names, the number of patients in each cluster, and the statistics about the most important clinical variables. The details of the remapping of the histories into the clusters are reported in the Appendix 5.

3.4. Careflow assessment and groups comparison

After having derived the nine clusters from administrative data only, it is interesting to clinically enrich them, in order to evaluate the so-called “temporal phenotypes”. In particular, the analysis of the relationships of those clusters with the patients’ outcomes may elucidate their clinical meaning and their capability of describing the evolution of the disease. In the following, we have considered three main endpoints: i) 10-year LRR-free survival probability; ii) 10-year DM-free survival probability and iii) 10-year overall survival probability.

Looking at LRR disease-free survival (Fig. 8), the group with the worse prognosis is represented by cluster 4b, i.e., second surgery after two months. As expected, these patients have a re-intervention that might be due to recurrence, which represents a severe clinical condition. The second group with short disease-free survival is group 6, neoadjuvant therapy. This group has a worse prognosis since it indirectly selects patients with advanced diseases, requiring neoadjuvant treatment. Kaplan-Meier analysis finds a statistically significant difference ($p < < 0.01$) between each group.

With regards to DM (Fig. 9), it is notable that cluster 6 (neoadjuvant) has the worst prognosis, followed by cluster 3, patients who undergo surgery, therapy and plastic surgery, and finally by cluster 2, patients with surgery and therapy. Again, Kaplan-Meier analysis confirmed statistically significant difference ($p < < 0.01$) between the groups.

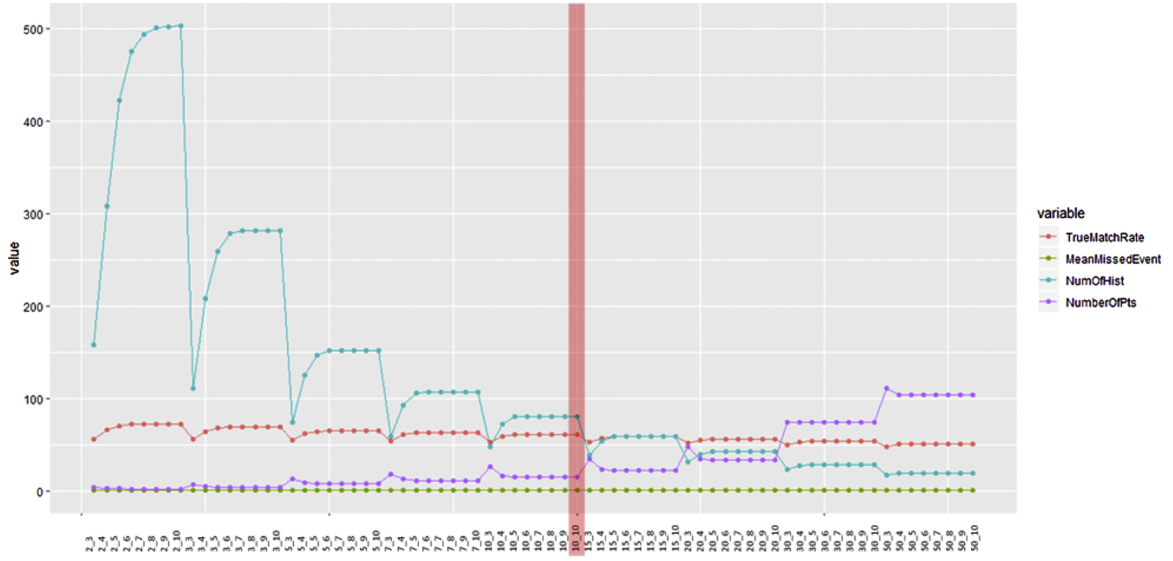


Fig. 6. Number of mined careflows (blue), average number of patients per careflow (purple), average number of missed events (green), and true match rate (red) at CFM parameters varying. The horizonal axis indicates the CFM parameters: min_support and max_length (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 4
List of the distinct events resulting after the application of the CFM step.

Events Label
Chemotherapy; biopsy
Chemotherapy; tumor
Day Hospital; SPU visit
Lumpectomy; prosthesis
Lumpectomy; quadrantectomy
Lumpectomy; removal
Mastectomy; prosthesis
Mastectomy; removal
Mastectomy; unilateral
Other exams and therapies; exams
Other exams and therapies; exercises
Other exams and therapies; tumor
Other exams and therapies; ultrasound
Plastic; prosthesis
Plastic; reconstruction
Skin graft / Lymph nodes operations; prosthesis
Skin graft / Lymph nodes operations; removal
Skin graft / Lymph nodes operations; skin

Finally, examination of overall survival (Fig. 10), revealed that cluster 6 is the group with the poorest prognosis. Cluster 9 also shows some deaths relatively close to the surgery, while clusters 3 and 2 are related to a lower median survival than all patients grouped together ($p < < 0.01$).

Given the results obtained by the Kaplan-Meier analyses, we investigated whether the clusters are significant predictors of survival if we consider also the available clinical variables in a statistical model. To this end, we have carried out a multivariate survival analysis by using Cox-Regressions to predict 10-year LRR-free, DM-free, and overall survival probabilities. Results in terms of Hazard Ratios (HR) are shown in Table 6.

It is possible to note that clusters are significant predictors of LRR, even when adjusting for clinical data. Only clusters 2 and 6 are predictors of DM- free survival, while none of them is an independent predictor of survival. Overall survival is indeed a complex function of different clinical and patient-related variables, and expectedly the cluster itself is not able to be independently predictive.

3.5. Comparison with baseline approach

In this paragraph, we compare the results presented in the previous section with the results obtained by running the CFM algorithm on events characterized only by the type of hospitalization (Admissions or SPU) and ward, obtained using a minimum support of 50 subjects and maximum history length of 10.

The clusters we obtained by applying this strategy are the following (see Supplementary Appendix 7 for their regrouping and outcomes comparisons):

- Cluster 0: Start with SPU
- Cluster 1: Admission in Breast Surgery and no other follow-ups
- Cluster 2: Multiple consecutive admissions in Breast Surgery
- Cluster 3: Admission in Breast Surgery followed by an Admission in Oncology
- Cluster 4: Admission in Breast Surgery and SPU in Breast Surgery
- Cluster 5: Admission in Breast Surgery and one or more SPUs in Oncology
- Cluster 6: Admission in Breast Surgery, SPU in Oncology and further Admission in Breast Surgery

The chord diagram in Fig. 11 represents the flows between clusters derived from the two different approaches. On the left, the results of the processing through TM and CFM are shown (P_), whereas the results of CFM only are represented on the right (AS_). Each cluster is represented by a fragment on the outer part of the circular layout, proportional to the number of subjects belonging to each group. The arcs, drawn between each fragment, display the flow of the same patients when associated with different clusters accordingly to the two strategies, TM and CFM coloured and only CFM in grey scale.

Table 7 shows a quantitative comparison of the clusters obtained with the two strategies. It is possible to observe that, while the most general histories (i.e. P_8, AS_1) almost perfectly map into each other, histories that we previously found significantly associated with increasing risks of adverse outcome (i.e. P_4 and P_6) were masked and fused together into more general and less meaningful ones (P_4 into AS_2, P_6 into AS_0).

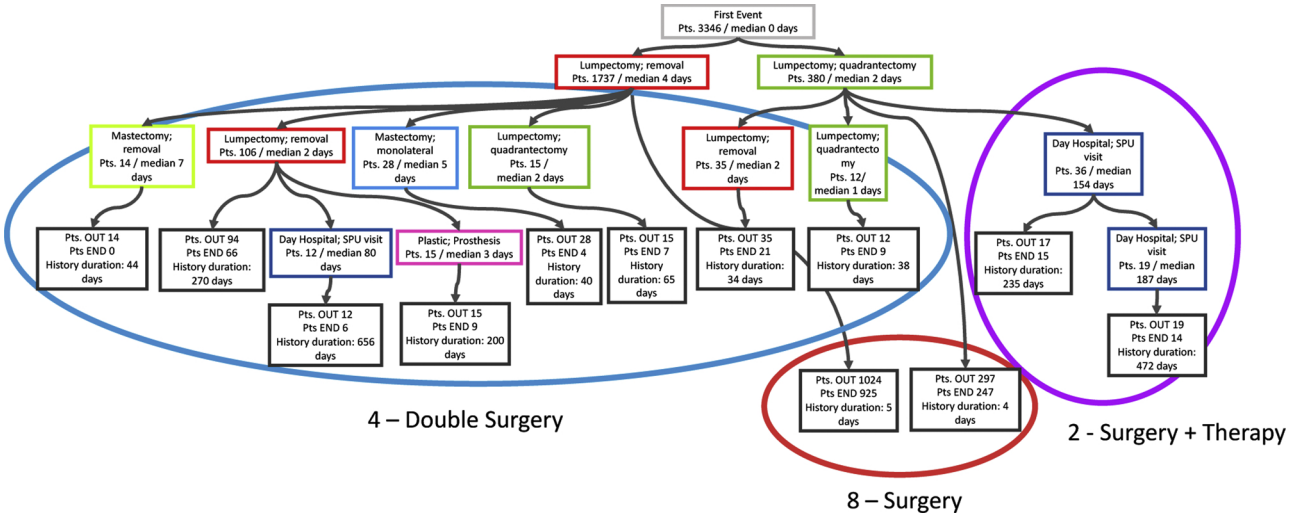


Fig. 7. Re-grouping of CFM results. The CFM results and the re-group of the branches as suggested by expert clinicians. It is possible to note that the regrouping follows the exact structure of the resulted DAG. Exit events report the number of subjects included in each branch. A patient careflow can end in an exit box for two reasons: (i) the patient has flown through all his/her events available in the data base (Pts.END), (ii) the patient has other events, but the careflow that would result from them has a minimum support lower than min_support or a length that is higher than max_length (Pts.OUT).

4. Discussion

In this paper we present an analysis pipeline that, starting from administrative data on breast cancer-related hospitalizations, is able to stratify the patients' population into meaningful groups that reflect the evolution of the disease in terms of clinical outcome.

Applying data and process mining to raw data collected in HIS might lead to poorly informative results due to the high variability that typically characterizes the patients' flows of care. For this reason, it is well-agreed that appropriate data preparation and preprocessing is crucial. One way to perform this step is to rely on domain knowledge. Nevertheless, a completely knowledge-driven approach might be highly expert-dependent and not easily reproducible. Using data-driven techniques to reduce the original variability is a valid alternative, and methodologies originating from the field of Natural Language Processing can be useful when the semantics are of importance. We have leveraged Topic Modeling to represent a single hospitalization with a topic that synthesizes the set of procedures carried out between admission and discharge on a specific patient. To identify the most frequent careflows enacted by the considered Breast Unit, we then ran a careflow mining algorithm on the sequence of hospitalizations represented by the corresponding topic.

As presented in some works in the literature [21,22], an alternative approach could have been to apply topic modeling to the entire set of ICD9-CM codes comprising the hospitalizations history of a patient, but this would have resulted into the loss of the information on the temporal order of events, which is instead preserved when using CFM. Recently, other works have dealt with the joint application of LDA and process mining for the automated discovery of clinical pathways in the case of intra-cerebral hemorrhage [9,10]. These papers focused on the optimization of the quality of the topics extracted by LDA through the specification of constraints to ensure that the same clinical activity performed on the same day would not be assigned to different topics, and to ensure that a single clinical activity would rank high in a limited number of topics. The process mining step, performed via the fuzzy miner algorithm [23], was limited to the reconstruction of clinical pathways that helped validating the LDA optimization process through the comparison of the obtained pathways to clinical guidelines.

While we took inspiration from these works to drive the LDA topic search on the basis of a set of indicators introduced in the Methods section, in this paper we focused on the interpretability and clinical significance of the extracted careflows, and on their potential to describe clinical phenotypes of patients. The CFM approach has several advantages with respect to the utilization of process mining algorithms in healthcare-related scenarios, such as the possibility to preserve the temporal order of events and to consider similar events occurring at different timestamps (e.g., two occurrences of breast surgery) as distinct. This is particularly important for histories such as the ones involving neoadjuvant treatments before the first breast surgery, which couldn't have been distinguished from careflows with treatment after surgery by using more standard process mining techniques. Nevertheless, one of the limitations of the approach is that it is not optimized for dealing with events occurring at the same time or overlapping. The introduction of the TM step prior to the use of CFM helps to mitigate this issue, as the events occurring together during the same hospitalization are managed and summarized by using the TM results. Another advantage of the proposed pipeline is related to the visualization of results as DAGs, which improve the results explicability and help in identifying of meaningful patients' subgroups.

We have previously applied the CFM algorithm to breast cancer data [5]. In that study, since the goal was to validate the methodology, we used simple events such as the type of stay (regular hospitalization or SPU visit) and the related ward. In addition, the information on the patients' outcome was not available, so it was not possible to compare the extracted careflows in terms of survival, but only to enrich them considering blood test results. In that work, we demonstrated that the algorithm was able to identify the most typical, high-level, patterns of care experienced by the patients included in the cohort, and to characterize them in terms of temporal and clinical information. In this work we report several advancements with respect to our previous research, furthermore we performed an evaluation of the proposed approach by comparing its results to the ones obtained by running the CFM algorithms on coarser clinical events represented, as in [5]. The results show that the new approach presented in this paper is able to provide finer phenotypes, significantly associated with relevant outcomes.

One of the main results of this study is that by using administrative data only, we have been able to identify clinically relevant trajectories able to stratify patients into informative groups with different evolution of the disease. As described in the Methods and in the Results section, we have considered two different representations of clinical events as input to the topic modeling step. The first one uses raw ICD9-CM codes, whereas the second one includes an initial knowledge-driven remapping step. Interestingly, the extracted topics were almost totally

Table 5
Cluster descriptive statistics. Age is reported as Mean (SD), all the other categorical variables are reposted as Number (% over the number of subjects in each cluster).

Cluster number Cluster name	1 Reconstruction/ Plastic surgery Only	2 Surgery + Therapy	3 Surgery + Therapy + Plastic Surgery	4a Double Surgery - Second Surgery Within 2 Months	4b Double Surgery - Second Surgery After 2 Months	5 Surgery + Plastic Surgery	6 Neoadjuvant	7 Surgery + Rehabilitation	8 Surgery	9 Surgery + Exams
Number of patients	284	668	114	132	110	285	53	42	1568	46
Age	49.7(11)	57(12.4)	51.2(10.9)	53.7(10.9)	61(13.2)	52.5(11)	55.2(11.7)	54.2(12.6)	63.3(13)	66.4(12.4)
Type of Surgery										
Lumpectomy	40(14.08)	504(75.45)	5(4.39)	119(90.15)	98(89.09)	73(25.61)	18(33.96)	25(59.52)	1389(88.58)	37(80.43)
Mastectomy	244(85.92)	164(24.55)	109(95.61)	13(9.85)	12(10.91)	212(74.39)	35(66.04)	17(40.48)	179(11.42)	9(19.57)
Histological Type										
DCIS In situ cancer	35(12.32)	9(1.35)	3(2.63)	23(17.42)	25(22.73)	47(16.49)	8(15.09)	2(4.76)	178(11.35)	6(13.04)
CDI Invasive ductal cancer	197(69.37)	561(83.98)	89(78.07)	87(65.91)	64(58.18)	183(64.21)	42(79.25)	31(73.81)	1134(72.32)	31(67.39)
CLI Invasive lobular cancer	52(18.31)	98(14.67)	22(19.3)	22(16.67)	21(19.09)	55(19.3)	3(5.66)	9(21.43)	256(16.33)	9(19.57)
Grading										
G1	22(7.75)	24(3.59)	3(2.63)	28(21.21)	20(18.18)	32(11.23)	2(3.77)	7(16.67)	275(17.54)	9(19.57)
G2	171(60.21)	272(40.72)	71(62.28)	71(53.79)	72(65.45)	176(61.75)	22(41.51)	24(57.14)	1001(63.84)	30(65.22)
G3	91(32.04)	372(55.69)	40(35.09)	33(25)	18(16.36)	77(27.02)	29(54.72)	11(26.19)	292(18.62)	7(15.22)
Neoadjuvant chemotherapy										
No	252(88.73)	596(89.22)	99(86.84)	130(98.48)	107(97.27)	256(89.82)	2(3.77)	37(88.1)	1422(90.69)	43(93.48)
Yes	30(10.56)	62(9.28)	13(11.4)	2(1.52)	2(1.82)	28(9.82)	51(96.23)	5(11.9)	127(8.1)	3(6.52)
Unknown	2(0.7)	10(1.5)	2(1.75)	0(0)	1(0.91)	1(0.35)	0(0)	0(0)	19(1.21)	0(0)
Cancer multifocality										
NO	194(68.31)	509(76.2)	65(57.02)	87(65.91)	83(75.45)	167(58.6)	38(71.7)	27(64.29)	1285(81.95)	38(82.61)
Yes	90(31.69)	159(23.8)	49(42.98)	45(34.09)	27(24.55)	118(41.4)	15(28.3)	15(35.71)	283(18.05)	8(17.39)
Staging										
Stage 0	41(14.44)	28(4.19)	7(6.14)	24(18.18)	24(21.82)	54(18.95)	31(58.49)	3(7.14)	216(13.78)	6(13.04)
Stage 1	106(37.32)	248(37.13)	30(26.32)	80(60.61)	62(56.36)	134(47.02)	8(15.09)	10(23.81)	948(60.46)	27(58.7)
Stage 2	90(31.69)	259(38.77)	45(39.47)	23(17.42)	22(20)	80(28.07)	7(13.21)	23(54.76)	334(21.3)	10(21.74)
Stage 3	47(16.55)	133(19.91)	32(28.07)	5(3.79)	2(1.82)	17(5.96)	7(13.21)	6(14.29)	70(4.46)	3(6.52)
Biomolecular subtype										
Luminal A	144(50.7)	170(25.45)	40(35.09)	73(55.3)	75(68.18)	172(60.35)	4(7.55)	26(61.9)	1070(68.24)	34(73.91)
Luminal B	94(33.1)	299(44.76)	55(48.25)	39(29.55)	24(21.82)	84(29.47)	25(47.17)	14(33.33)	382(24.36)	9(19.57)
Her2+	20(7.04)	79(11.83)	7(6.14)	5(3.79)	3(2.73)	6(2.11)	22(41.51)	1(2.38)	27(1.72)	0(0)
TNBC	26(9.15)	120(17.96)	12(10.53)	15(11.36)	8(7.27)	23(8.07)	2(3.77)	1(2.38)	89(5.68)	3(6.52)
Hormone therapy										
NO	52(18.31)	239(35.78)	25(21.93)	32(24.24)	31(28.18)	50(17.54)	25(47.17)	5(11.9)	285(18.18)	11(23.91)
Yes	232(81.69)	429(64.22)	89(78.07)	100(75.76)	79(71.82)	235(82.46)	28(52.83)	37(88.1)	1283(81.82)	35(76.09)
Chemotherapy										
NO	144(50.7)	102(15.27)	15(13.16)	100(75.76)	101(91.82)	212(74.39)	0(0)	23(54.76)	1334(85.08)	42(91.3)
Yes	140(49.3)	566(84.73)	99(86.84)	32(24.24)	9(8.18)	73(25.61)	53(100)	19(45.24)	234(14.92)	4(8.7)
Radiotherapy										
NO	176(61.97)	125(18.71)	79(69.3)	41(31.06)	37(33.64)	200(70.18)	25(47.17)	16(38.1)	350(22.32)	20(43.48)
Yes	108(38.03)	543(81.29)	35(30.7)	91(68.94)	73(66.36)	85(29.82)	28(52.83)	26(61.9)	1218(77.68)	26(56.52)
Lipofilling intervention										
No	119(41.9)	603(90.27)	89(78.07)	118(89.39)	102(92.73)	199(69.82)	50(94.34)	36(85.71)	1525(97.26)	42(91.3)
Yes	165(58.1)	65(9.73)	25(21.93)	14(10.61)	8(7.27)	86(30.18)	3(5.66)	6(14.29)	43(2.74)	4(8.7)
Local/loco-regional recurrence (Yes)	12(4.23)	50(7.49)	11(9.65)	8(6.06)	39(35.45)	25(8.77)	5(9.43)	2(4.76)	32(2.04)	7(15.22)
Metastasis - Distant recurrence (Yes)	22(7.75)	116(17.37)	22(19.3)	3(2.27)	7(6.36)	15(5.26)	10(18.87)	5(11.9)	48(3.06)	2(4.35)
Exitus (Yes)	16(5.63)	66(9.88)	14(12.28)	3(2.27)	5(4.55)	7(2.46)	5(9.43)	4(9.52)	46(2.93)	6(13.04)

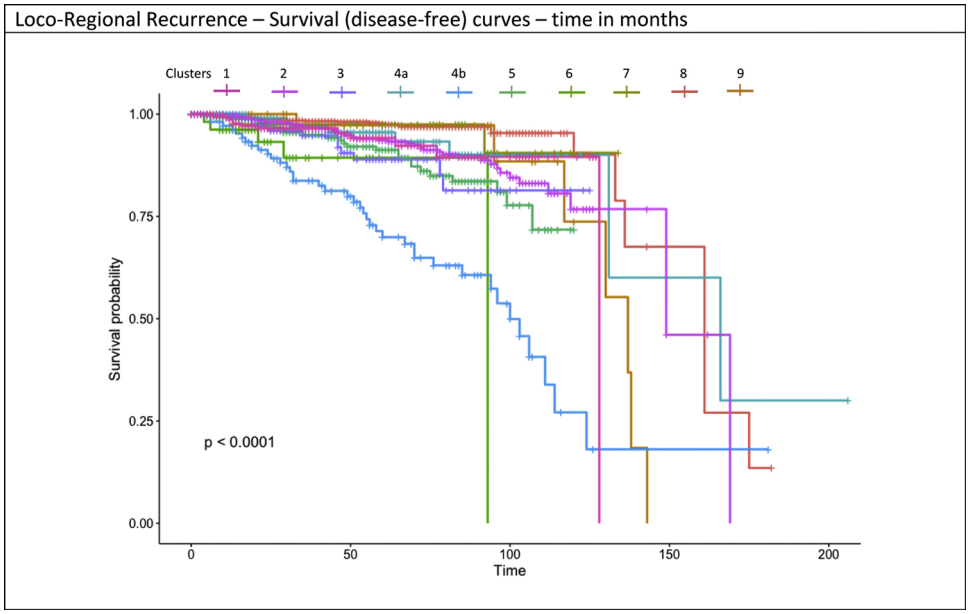


Fig. 8. Local-regional recurrence survival (disease) free curves in the Clusters – Time in months.

overlapping. This indicates that TM is able to extract the needed information independently from the type of coding used for the events and from the representation language (the ICD9-CM description was in Italian, while the remapping was in English). On the basis of this result, we argue that in the data and clinical domain we investigated, using raw ICD9-CM codes is preferable, as it avoids an additional expert-based step, which could limit the reproducibility of the results. However, this might not be true in other domains or data sources, and this merits further investigation.

While the initial knowledge-based recoding has been shown to be unnecessary, some other steps needed the close interaction among experts to properly guide the algorithms and interpret the results [24]. In particular, the expert-based validation of the label assigned to each topic is extremely important to capture the clinical significance of the extracted words, and the post-processing of the results of CFM to identify careflows that represent the same trajectory is crucial to properly identify the patients' phenotypes. An example of this second step is related to Group 8, which includes patients who had only undergone surgery at ICSM. In this case, we merged into a single group those patients who underwent a lumpectomy, patients who underwent a mastectomy, and patients who underwent lumpectomy together with operations on the lymph nodes.

When process mining approaches are applied to the healthcare domain, especially to secondary data coming from clinical routine processes, a set of domain related, expert-based interventions are usually needed. These manual interventions allow avoiding the spaghetti-like models often resulting from fully automated approaches, and can be performed either (a) by pre-processing data via the initial formalization of the domain knowledge, for example with ontology-based approaches [25,26], or (b) with unstructured data and post-processing tailored interventions, as in the case presented in this paper. These interventions have the potential to produce more detailed process

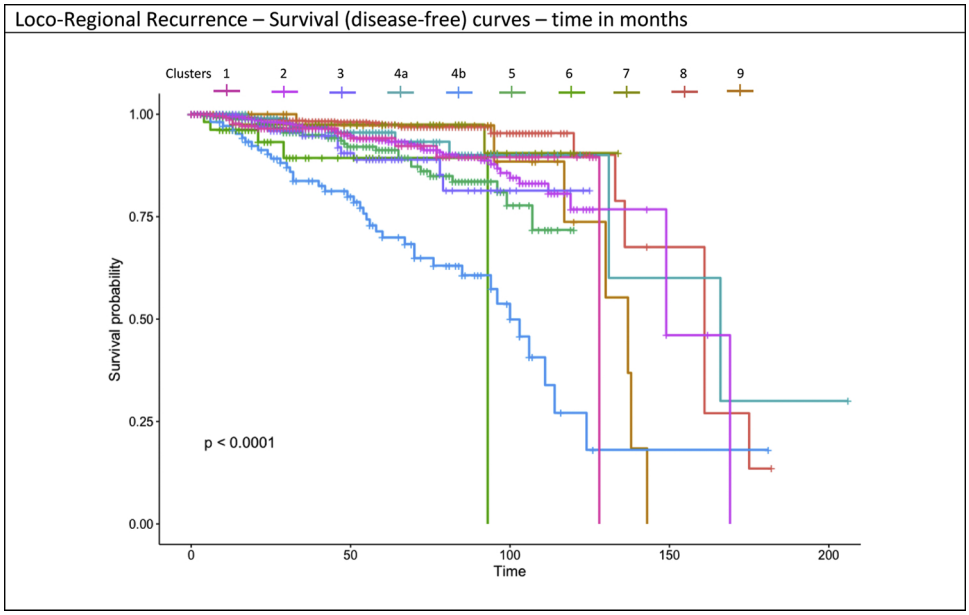


Fig. 9. Distant metastasis survival (disease) free curves in the Clusters – Time in months.

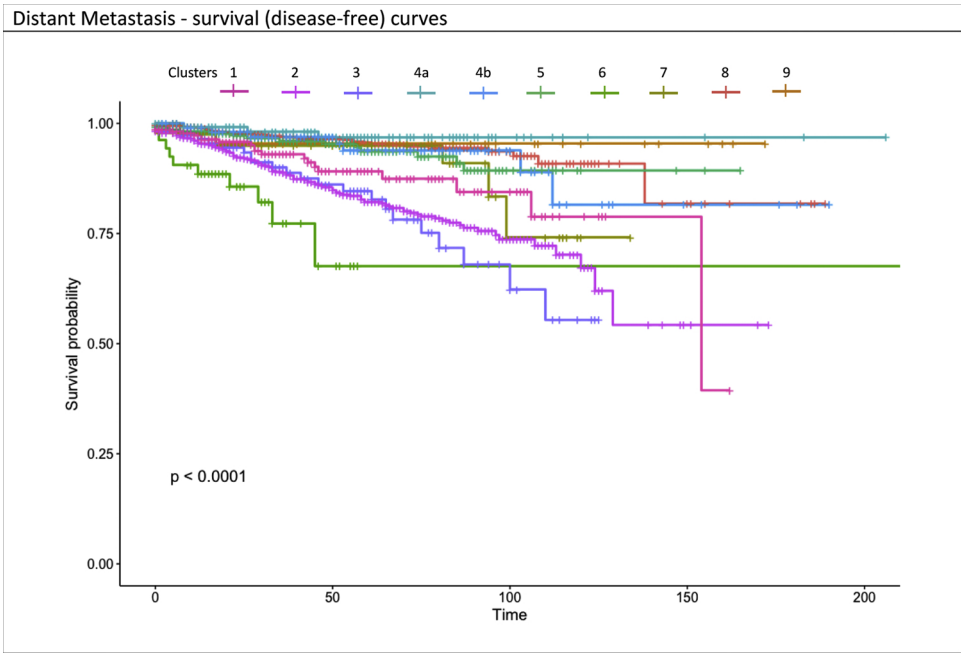


Fig. 10. Survival (registered deaths) in the Clusters – Time in months.

models than completely automatized methods, especially when analysing complex clinical processes data with unstructured components, thus providing clinicians with readable results – also including information about the specific organizational set-ups of the hospital.

The results we have obtained are meaningful with respect to the Breast Unit we studied. In particular, we have been able to distinguish those patients only undergoing surgery at the hospital from the patients who instead carry out the entire process of care within the studied hospital. Among these, we mainly identified three groups: (i) patients who undergo breast surgery possibly followed by plastic reconstruction and then chemotherapy, (ii) patients who undergo multiple surgeries due to cancer recurrence, and (iii) patients who undergo neoadjuvant therapy before surgery. Such groups turned out to be different in terms of clinical endpoints, considered as onset of metastases, local recurrences, and overall survival.

Once these sub-groups are extracted and clinically validated, they might constitute the basis for automatic case retrieval in more complex architectures [27]. In particular, it would be both possible to query the

Table 6
Cox Regression results reported as Hazard Ratio + 95 % CI for HR and significance codes: '***' 0.001 '**' 0.01 '*' 0.05 (p-values).

		LRR	*DM*	*Death*
Cluster	**1 - Only Reconstruction/ Plastic surgery**	3.06 (1.53–6.11) **	n.s.	n.s.
	2 - Surgery + Therapy	1.87 (1.15–3.04)*	1.97 (1.29–2.99)**	n.s.
	3 - Surgery + Therapy + Plastic Surgery	3.23 (1.51–6.91)**	n.s.	n.s.
	4 - Double Surgery	–	–	–
	4a - Double Surgery - Second Surgery Within 2 Months	11.59 (7.16–18.76)***	n.s.	n.s.
	4b - Double Surgery - Second Surgery After 2 Months	n.s.	n.s.	n.s.
	5 - Surgery + Plastic Surgery	4.52 (2.64–7.75)***	n.s.	n.s.
	6 - Neoadiuvant	3.33 (1.16–9.56)*	2.43 (1.14–5.19)*	n.s.
	7 - Surgery + Rehabilitation	n.s.	n.s.	n.s.
	8 - Surgery	Reference	Reference	Reference
	9 - Surgery–+ Exams	2.83 (1.21–6.62)*	n.s.	n.s.
Age		1.01(1–1.02)***	n.s.	1.02(1.01–1.03)**
Type of Surgery	**Lumpectomy**	Reference	Reference	Reference
	Mastectomy	n.s.	1.54(1.11–2.13)**	n.s.
Staging	**Stage 0**	Reference	Reference	Reference
	Stage 1	n.s.	n.s.	n.s.
	Stage 2	n.s.	2.24(1.28–3.92)**	3.44(1.75–6.76)***
	Stage 3	n.s.	4.60(2.58–8.13)***	6.72(3.35–13.46)***
Grading	**G1**	Reference	Reference	Reference
	G2	n.s.	2.08(1.04–4.16)*	3.75(1.3–9.84)*
	G3	n.s.	2.40(1.18–4.89)*	4.15(1.44–11.92)*
Hormone therapy	**Yes**	0.49(0.33–0.75)***	n.s.	n.s.
Chemotherapy	**Yes**	n.s.	n.s.	n.s.
Radiotherapy	**Yes**	n.s.	n.s.	0.63(0.46–0.87)**
Neoadjuvant chemotherapy	**Yes**	2.56(1.69–3.88)***	2.82(1.99–3.98)***	3.40(2.42–4.78)***
Biomolecular subtype	**Luminal A**	Reference	Reference	Reference
	Luminal B	1.45(1.01–2.09)*	n.s.	1.46(1.01–2.11)*
	Her2+	n.s.	n.s.	n.s.
	TNBC	n.s.	n.s.	2.21(1.3–3.77)**
Lipofilling intervention	**Yes**	n.s.	0.62(0.42–0.92)*	n.s.

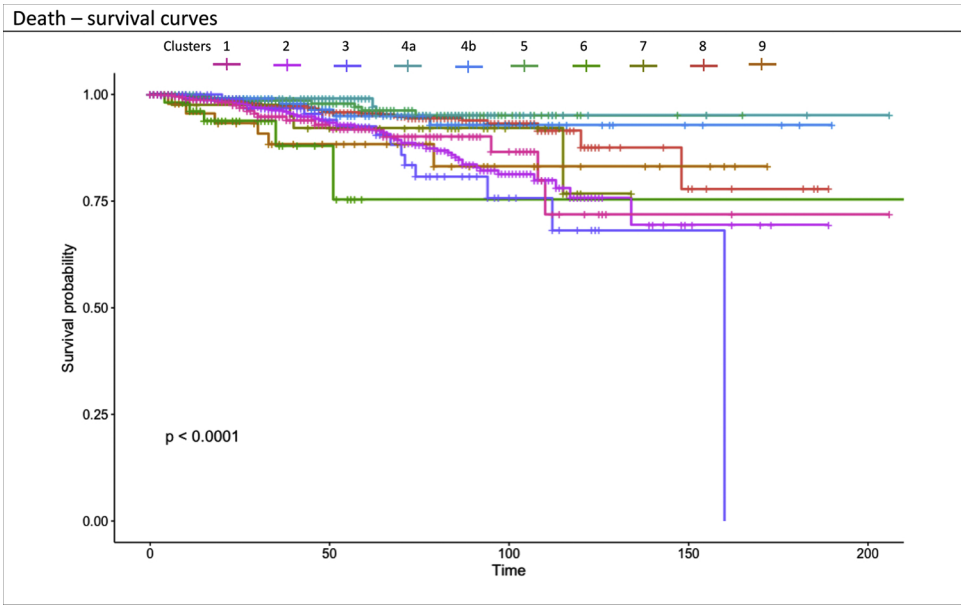


Fig. 11. Chord diagram visualizing the relations between clusters derived via TM and CFM on the basis of procedures (P_ on the left) and clusters derived from admission and SPU (AS_ on the right).

architecture on the basis of the specific subgroups, and, once a new patient is considered, it would be possible to assign him to the trajectory that is more similar according to a suitable function [4]. Such an approach could be suitably exploited to rapidly analyze large datasets in order to derive clinical management strategies that could be related to improved or decreased long-term oncologic outcomes.

The proposed analyses have some limitations. Some limitations are inherently derived from the application of process mining technologies to healthcare pathways analysis, and are related to challenges of applying these approaches to unstructured processes, the lack of established data reference models and the necessity to implement custom solutions for each case study [28]. More specifically, when the context data are derived from administrative systems or healthcare logistic systems, on the one hand the number of events to be mined for each patient can be low on average when considering coarse granularities such as an entire hospitalization, on the other hand using finer granularity will produce unreadable spaghetti-like processes. Previous works [29] identified these limitations and highlighted how in healthcare systems events granularity is often too low for process mining algorithms to identifying the correct control-flow as the ordering of events.

Some limitations are specific to this work. First of all, we considered only semi-structured data, corresponding to the description of the ICD9-CM codes included in discharge letters. While the process of care for breast cancer patients develops mainly through events with a structured discharge letter prepared for billing purposes, this might not be the case for other diseases. To complete the view on the clinical history of the patient, it would be possible to add to the analysis textual reports released during outpatient visits. Other relevant information that might be worth exploring is related to comorbidities, which can have an effect on both the mined careflows and the results, thus acting as confounders.

In this paper we have considered the procedures performed in the same hospitalization as a whole, regardless of their temporal order. This is possible as hospitalizations for breast cancer are relatively short-term and, even more importantly, usually include procedures with a specific

Table 7
Adjacency matrix of the mapping between clusters derived via TM and CFM on the basis of procedures (P_ rows) and clusters derived from admission and SPU (AS_ columns). The table reports numbers and percentages calculated over the number of subjects for rows.

	AS_0	AS_1	AS_2	AS_3	AS_4	AS_5	AS_6
P_1	NA	62 (21.83%)	**170 (59.85%)**	3 (1.06%)	15 (5.28%)	22 (7.74%)	12 (4.22%)
P_2	6 (0.89%)	35 (5.23%)	NA	11 (1.64%)	NA	**514 (76.94%)**	102 (15.26%)
P_3	NA	4 (3.5%)	11 (9.65%)	NA	NA	2 (1.75%)	**97 (85.08%)**
P_4	39 (16.11%)	3 (1.24%)	**161 (66.53%)**	NA	39 (16.11%)	NA	NA
P_5	2 (0.70%)	10 (3.51%)	**270 (94.74%)**	NA	3 (1.05%)	NA	NA
P_6	**53 (100%)**	NA	NA	NA	NA	NA	NA
P_7	NA	6 (14.28%)	NA	NA	NA	**22 (52.38%)**	14 (33.34%)
P_8	47 (2.99%)	**1368 (87.24%)**	80 (5.10%)	27 (1.72%)	8 (0.51%)	31 (1.98%)	7 (0.45%)
P_9	NA	**23 (50%)**	NA	10 (21.74%)	NA	5 (10.87%)	8 (17.39%)

	AS_0	AS_1	AS_2	AS_3	AS_4	AS_5	AS_6
P_1	NA	62 (21.83%)	**170 (59.85%)**	3 (1.06%)	15 (5.28%)	22 (7.74%)	12 (4.22%)
P_2	6 (0.89%)	35 (5.23%)	NA	11 (1.64%)	NA	**514 (76.94%)**	102 (15.26%26 %)
P_3	NA	4 (3.5 %)	11 (9.65 %)	NA	NA	2 (1.75 %)	**97 (85.08 %)**
P_4	39 (16.11 %)	3 (1.24 %)	**161 (66.53 %)**	NA	39 (16.11 %)	NA	NA
P_5	2 (0.70 %)	10 (3.51 %)	**270 (94.74 %)**	NA	3 (1.05 %)	NA	NA
P_6	**53 (100 %)**	NA	NA	NA	NA	NA	NA
P_7	NA	6 (14.28 %)	NA	NA	NA	**22 (52.38 %)**	14 (33.34 %)
P_8	47 (2.99 %)	**1368 (87.24 %)**	80 (5.10 %)	27 (1.72 %)	8 (0.51 %)	31 (1.98 %)	7 (0.45 %)
P_9	NA	**23 (50 %)**	NA	10 (21.74 %)	NA	5 (10.87 %)	8 (17.39 %)

overall goal, resulting in topics with high inner coherence. For other applications this might not be the case, and it could be necessary to change the temporal granularity, by considering for example the single hospitalization day, as it has been done in other works [9].

Finally, the CFM algorithm discovers careflows that include patients who undergo the same temporal sequence of events. In some cases, even though the sequence is the same, the temporal gap between two or more events could discriminate patients with different outcomes. In this work, we faced this problem in the case of Cluster 4 (Double surgery), where the time between the first and the second intervention is important to distinguish among re-interventions that are complementary to the first surgery or related to a change in the condition of the patient instead (e.g. a recurrence). At the moment, this step was expert-driven and manually performed after checking the CFM results, potentially generating a non-uniformity in results interpretation. In the future, a constraint on the temporal duration of events and transitions could be included in the CFM search strategy.

Nevertheless, the approach presented in this paper has the potential to help clinicians and hospital decision makers to exploit routinely collected administrative data to have a snapshot of the population of patients they are treating. This would allow identifying groups of critical patients or hidden care patterns, which would need further attention and require to plan specific interventions.

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.artmed.2020.101855.

References

[1] AIOM. I numeri del cancro in Italia [Cancer numbers in Italy]. 2018http://www.registri-tumori.it/PDF/AIOM2014/I_numeri_del_cancro_2014.pdf.

[2] Canavese G, Del Mastro Lucia, Frassoldati A, Montemurro F, Puglisi F, Mimma RSG. Linee guida neoplasia della mammella 2017. Aiom 2017.

[3] Baker K, Dunwoodie E, Jones RG, et al. Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. Int J Med Inform 2017. https://doi.org/10.1016/j.ijmedinf.2017.03.011.

[4] Dagliati A, Tibollo V, Cogni G, et al. Careflow mining techniques to explore type 2 diabetes evolution. J Diabetes Sci Technol 2018. https://doi.org/10.1177/1932296818761751.

[5] Dagliati A, Sacchi L, Zambelli A, et al. Temporal electronic phenotyping by mining careflows of breast cancer patients. J Biomed Inform 2017. https://doi.org/10.1016/j.jbi.2016.12.012.

[6] van der Aalst W. Process mining: discovering and improving Spaghetti and Lasagna processes. 2011 IEEE Symp Comput Intell Data Min 2011:1–7. https://doi.org/10.1109/CIDM.2011.6129461.

[7] Kaymak U, Mans R, Van De Steeg T, et al. On process mining in health care. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics 2012. https://doi.org/10.1109/ICSMC.2012.6378009.

[8] Liu S, Hauskrecht M. Nonparametric regressive point processes based on conditional gaussian processes. 2019.

[9] Xu X, Jin T, Wei Z, et al. Incorporating topic assignment constraint and topic correlation limitation into clinical goal discovering for clinical pathway mining. J Healthc Eng 2017. https://doi.org/10.1155/2017/5208072.

[10] Xu X, Jin T, Wei Z, et al. TCPM: topic-based clinical pathway mining. Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016 2016. https://doi.org/10.1109/CHASE.2016.17.

[11] Banziger R, Basukoski A, Chaussalet T. Discovering business processes in CRM systems by leveraging unstructured text data. Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2018 2019. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00257.

[12] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003.

[13] Sorrentino L, Regolo L, Scoccia E, et al. Autologous fat transfer after breast cancer surgery: an exact-matching study on the long-term oncological safety. Eur J Surg Oncol 2019. https://doi.org/10.1016/j.ejso.2019.05.013.

[14] Pentaho - Data Integration and Analytics Platform | Hitachi Vantara.

[15] R: The R Project for Statistical Computing.

[16] Blei D, Carin L, Dunson D. Probabilistic topic models. IEEE Signal Process Mag 2010. https://doi.org/10.1109/MSP.2010.938079.

[17] Grün B, Hornik KR. Package 'topicmodels'. CRAN; 2017.

[18] Maier D, Waldherr A, Miltner P, et al. Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun Methods Meas 2018. https://doi.org/10.1080/19312458.2018.1430754.

[19] Gatta R, Lenkowicz J, Vallati M, et al. pMineR: an innovative R library for performing process mining in medicine. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2017. https://doi.org/10.1007/978-3-319-59758-4_42.

[20] Elixhauser A, Steiner C, Palmer L, Clinical Classifications Software (CCS). US Agency Healthc Res Qual 2014. 2014. p. 1–54http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp.

[21] Huang Z, Dong W, Ji L, et al. Discovery of clinical pathway patterns from event logs using probabilistic topic models. J Biomed Inform 2014;47:39–57. https://doi.org/10.1016/j.jbi.2013.09.003.

[22] Huang Z, Lu X, Duan H. On mining clinical pathway patterns from medical behaviors. Artif Intell Med 2012;56:35–50. https://doi.org/10.1016/j.artmed.2012.06.002.

[23] Günther CW, Van Der Aalst WMP. Fuzzy mining - adaptive process simplification based on multi-perspective metrics. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2007.

[24] Van Der Aalst WMP. Process mining. Process Min 2011;5:301–17. https://doi.org/10.1007/978-3-642-19345-3.

[25] Pereira Detro S, Santos EAP, Panetto H, et al. Applying process mining and semantic reasoning for process model customisation in healthcare. Enterp Model Inf Syst Archit 2019. https://doi.org/10.1080/17517575.2019.1632382.

[26] Leonardi G, Striani M, Quaglini S, et al. Leveraging semantic labels for multi-level abstraction in medical process mining and trace comparison. J Biomed Inform 2018. https://doi.org/10.1016/j.jbi.2018.05.012.

[27] Lamy JB, Sekar B, Guezennec G, et al. Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. Artif Intell Med 2019. https://doi.org/10.1016/j.artmed.2019.01.001.

[28] Rojas E, Munoz-Gama J, Sepulveda M, et al. Process mining in healthcare: a literature review. J Biomed Inform 2016;61:224–36. https://doi.org/10.1016/j.jbi.2016.04.007.

[29] Mans RS, Van Der Aalst WMP, Vanwersch RJB, et al. Process mining in healthcare: data challenges when answering frequently posed questions. Methods Mol Biol 2015:1246. https://doi.org/10.1007/978-3-642-36438-9_10.