Data-Driven Treatment Pathways Mining for Early Breast Cancer Using cSPADE Algorithm and System Clustering

Qing Yang
Sichuan University West China Hospital
Ting Luo
Sichuan University West China Hospital
Wei Zhang (weizhanghx@163.com)
Sichuan University West China Hospital https://orcid.org/0000-0003-3113-9577
Xiaorong Zhong
Sichuan University West China Hospital
Ping He
Sichuan University West China Hospital
Hong Zheng
Sichuan University West China Hospital

Research article

Keywords: Breast cancer, Data mining, Clinical pathway, Sequential pattern mining, Cluster analytics

Posted Date: September 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-69976/v1

Version of Record: A version of this preprint was published at The International Journal of Health Planning and Management on April 20th, 2022. See the published version at https://doi.org/10.1002/hpm.3483.

Abstract

Background: Due to the multidimensional, multilayered, and chronological order of the cancer data in this study, it was challenging for us to extract treatment paths. Therefore, it was necessary to design a new data mining scheme to effectively extract the treatment path of breast cancer. To determine whether the cSPADE algorithm and system clustering proposed in this study can effectively identify the treatment pathways for early breast cancer.

Methods: We applied data mining technology to the electronic medical records of 6891 early breast cancer patients to mine treatment pathways. We provided a method of extracting data from EMR and performed three-stage mining: determining the treatment stage through the cSPADE algorithm → system clustering for treatment plan extraction → cSPADE mining sequence pattern for treatment. The Kolmogorov-Smirnov test and correlation analysis were used to cross-validate the sequence rules of early breast cancer treatment pathways.

Results: We unearthed 55 sequence rules for early breast cancer treatment, 3 preoperative neoadjuvant chemotherapy regimens, 3 postoperative chemotherapy regimens, and 2 chemotherapy regimens for patients without surgery. Through 5-fold cross-validation, Pearson and Spearman correlation tests were performed. At the significance level of $P < 0.05$, all correlation coefficients of support, confidence and lift were greater than 0.89. Using the Kolmogorov-Smirnov test, we found no significant differences between the sequence distributions.

Conclusions: The cSPADE algorithm combined with system clustering can achieve hierarchical and vertical mining of breast cancer treatment models. By uncovering the treatment pathways of early breast cancer patients by this method, the real-world breast cancer treatment behavior model can be evaluated, and it can provide a reference for the redesign and optimization of the treatment pathways.

Background

At present, breast cancer ranks first among the diseases affecting women in the world and is the leading cause of cancer deaths [1]. Faced with the high incidence of cancer, how to reduce mortality and prolong survival has become a hot topic of research. To this end, a variety of treatment methods have been developed, including surgical treatment, radiotherapy, chemotherapy, endocrine therapy and targeted therapy, and breast cancer treatment has begun to develop into personalized and precise treatments [2, 3]. How to effectively use these treatments to improve the medical effect and quality is a problem we face.

However, the current level of breast cancer diagnosis and treatment in China lags behind that in European and North American countries. There are obvious gaps in the level of diagnosis and treatment equipment and technology in different regions and different hospitals. Inadequate compliance and lack of standardization in guidelines, specifications and paths have hindered the improvement of the overall level of care[4]. However, Chinese women are different from women in Europe and the United States in many aspects, such as breast structure, sex hormone levels, eating habits, and living environment. Foreign guidelines and clinical pathways may not be completely applicable [5]. However, building a new clinical pathway from scratch is a time-consuming task for medical staff because it involves aspects such as multidisciplinary collaboration, sequential design, and outcome measurement [6]. Regarding the construction of the treatment path, traditional path development requires multiple experienced doctors, nurses and other related personnel to spend much time collecting data, ensure that the data are fully evidence-based, and discuss together[6]. Differences in opinions often result from differences in personal experience, leading to deviations between the established path and the actual operation. There are some defects in the quality of clinical pathway design[7].

In recent years, due to the development of informatics, software engineering, mathematics and other disciplines, interdisciplinary research has begun to emerge, and treatment pathway construction methods based on process mining

technology have appeared [8–11]. The development of electronic medical records (EMR) provides the possibility for the extraction and optimization of treatment paths [12–14]. Most process mining algorithms can automatically build process patterns, which are very suitable for understanding and can be used for process optimization [15–17]. Recently, many studies have focused on developing sequential pattern mining methods to discover real-world treatment behavior patterns from clinical data, which has become a research hotspot [18–20]. However, current research focuses on the mining of drug treatment models [21–23].

Due to the multidimensional, multilayered, and chronological order of the cancer data in this study, it was challenging for us to extract treatment paths. Therefore, it was necessary to design a new data mining scheme to effectively extract the treatment path of breast cancer. Sequence data consist of a series of ordered elements or events and may not include specific time concepts, such as customer shopping sequences, website click streams and biological sequences. This type of data does not process data at one point in time but rather at a large number of points in time. The sequence mode seeks to find the order between these data items. At present, sequential pattern mining has been applied to flood alarms[24], gene regulatory sequences[25], electronic health records (EHR) workflow[26], atrial fibrillation treatment path[27], and disease comorbidity pattern [28].

This study pioneered the joint use of two algorithms emerging in the field of data mining: the cSPADE algorithm + system clustering. Based on this, a three-stage mining method is proposed: the treatment phase is determined by the cSPADE algorithm → system clustering for treatment plan extraction → cSPADE mining sequence mode for treatment. This method realizes the extraction of breast cancer treatment pathways, which can be used to evaluate real-world breast cancer treatments, and provides a reference for the redesign and optimization of treatment pathways.

Methods

Materials

West China Hospital of Sichuan University is one of the top general hospitals in China and receives more than 260,000 inpatients from all over the country and internationally every year. We extracted data for 6891 stage 0–III breast cancer patients from West China Hospital of Sichuan University from 2011 to 2017. The average age of the patients was 48.67 ± 10.41 years. According to the patient's registration number, longitudinal tracking was performed. There were 41,070 inpatient medical records, 381,830 outpatient medical records, and more than 10 million doctor's orders. We extracted general information and clinical characteristics of patients, diagnosis, admission and discharge dates, and all inpatient and outpatient orders. The data used in this study are anonymous. Although according to Chinese law, this retrospective EMR study does not require the ethical approval of the regional ethics review committee, we still applied for and obtained the approval of the ethics committee of West China Hospital of Sichuan University (approval number: 2017 − 255).

Data Preparation

We labeled medical orders related to surgery, radiotherapy, chemotherapy, targeted therapy, and endocrine therapy. Fifty-nine of 6891 breast cancer patients did not undergo primary antitumor treatment; that is, 6,832 patients ultimately received primary treatment. There are 5758 types of doctor orders. Orders not related to breast cancer surgery, chemotherapy, radiotherapy, endocrine therapy, or targeted therapy were excluded, such as primary care or saline. Then, 138 original medical order names related to surgery, radiotherapy, chemotherapy, endocrine therapy, and targeted therapy were marked.

Mining the treatment pathways of early breast cancer based on the cSPADE algorithm and systematic clustering

The data in this study exist in multiple dimensions: doctor's orders, electronic medical records, outpatient records and other dimensions of data. Moreover, there are multiple levels of data: the first level of breast cancer treatment includes surgery, radiotherapy, chemotherapy, targeted therapy, etc. The second level includes different treatment options, such as surgery including radical surgery, modified radical surgery, breast-conserving surgery, and breast reconstruction, which follow a chronological order, such as chemotherapy first → surgery, or surgery first → chemotherapy. There are no methods for addressing such complicated path mining in the literature regarding the treatment of cancer patients. This study pioneered the joint use of two algorithms emerging in the field of data mining: the cSPADE algorithm + system clustering. The cSPADE algorithm was used to mine sequential patterns of treatment paths with time sequence, and then, system clustering was used to achieve dimensionality reduction of different treatment methods.

The three-stage mining method we proposed is as follows: determining the treatment stage through the cSPADE algorithm → systematic clustering for treatment plan extraction → cSPADE mining of the sequential mode of treatment. After the first analysis, we identified different treatment stages. The second step was to combine different study findings and clinicians' suggestions for each stage or use cluster analysis to summarize the typical treatment plan. The third step was to use sequential pattern mining to link treatment plans in different treatment stages in chronological order, find corresponding rules, and finally determine the main treatment path. The process of data mining is shown in Fig. 1.

<Fig. 1 about here>

Figure 1 Mining process of early breast cancer treatment

Identifying all frequent sequential patterns in a transaction database, such as in a large EMR database, requires efficient algorithms to handle large search spaces, and many different algorithms have been developed. In 2001, Zaki described an algorithm called sequential pattern discovery using equivalence classes (SPADE) that uses many strategies to make sequential pattern mining more efficient [29]. Sequential pattern mining usually starts with a transaction database, where each transaction has three fields: a "sequence" corresponding to the subject of the sequence (this study is the patient's medical record number); "transaction time" (the timing of the doctor's order in this study); and "transaction-related items" (the doctor's order for this study).

SPADE starts with a horizontal database layout, as shown in Table 1, but it then, converts the dataset into vertical "id lists" for each item, each item containing all sequence IDs and transaction times. Storing a vertical id list allows us to find sequential patterns using the intersections of the id lists. For example, the intersection of an id list of two items can be used to find sequential patterns (unilateral mastectomy, exemestane tablets). This method minimizes the number of database scans required. SPADE also takes advantage of common prefixes between sequences to reduce memory requirements. cSPADE is a version of SPADE that contains constraints on sequences.

Table 1
Examples of transaction databases

Sequence-id	Transaction-time	Items
30051	2017-08-24	single lumpectomy, arbitrary flap repair
30051	2018-03-02	tamoxifen citrate tablets
34776	2015-09-14	unilateral mastectomy
34776	2016-01-02	exemestane

<Table 1 about here>

Table 1 Examples of transaction databases

Mining Primary Breast Cancer Treatment Pathways

A total of 6,832 patients underwent primary treatment in this study. The main treatments for early breast cancer include surgery, chemotherapy, radiotherapy, endocrine therapy, and targeted therapy. In this study, the cSPADE algorithm was used to mine sequence patterns for early breast cancer treatment, and the sequence pattern of treatment was used as the primary treatment pathways for early breast cancer. We set the support to 0.15 and the maximum length to 10.

Extraction of the breast cancer treatment plan

In the first step of data mining, we identified the primary treatment path for early breast cancer and identified the stage of treatment for early breast cancer: preoperative stage (neoadjuvant chemotherapy) → surgery → postoperative chemotherapy → radiotherapy → endocrine therapy. To further analyze the different treatment plans, we further subdivided the treatment plans.

Breast cancer surgery methods are divided into expanded radical surgery, radical surgery, modified radical surgery, breast-conserving surgery, breast reconstruction, sentinel lymph node biopsy, and supraclavicular lymph node dissection.

Because chemotherapy may occur before, after, or during the treatment of patients who have not undergone surgery, we distinguish patients with chemotherapy orders based on the time of surgery and analyze the chemotherapy regimen in three time periods. Because multiple drugs were used in the same chemotherapy process, based on the co-occurrence of drugs, we clustered the orders of preoperative, postoperative, and nonoperative chemotherapy separately.

After referring to relevant literature and consulting breast cancer radiotherapy experts, we did not subdivide the radiotherapy plan. Endocrine drugs were classified into aromatase inhibitors, selective estrogen receptor modulators, and progestins according to their original categories.

Mining Of Secondary Treatment Pathways For Early Breast Cancer

After identifying the different treatment stages for the primary treatment pathway, we subdivided the treatment plan, and based on this, we examined the secondary pathways for early breast cancer treatment. We continued to use R language and the cSPADE algorithm, setting support = 0.02, confidence = 0.5, lift = 1. After generating sequence rules, functions were used to remove redundant rules.

Cross-validation

To verify the stability and accuracy of the results, we also used 5-fold cross validation. The Kolmogorov-Smirnov test, Pearson correlation analysis and Spearman correlation analysis were used to cross-validate the sequence rules of early breast cancer treatment pathways.

Results

Overall situation of early breast cancer treatment

Figure 2 shows that from 2011 to 2017, the utilization rate of surgery, radiotherapy, endocrine therapy, and targeted therapy has been increasing, and the utilization rate of chemotherapy has been decreasing. Among them, the increase in radiotherapy and targeted therapy is very obvious, with the utilization rate of radiotherapy increasing from 30.8–42.5% and the utilization rate of targeted therapy increasing from 5.3–18.3%.

<Fig. 2 about here>

Figure 2 Trends in the use of different treatments for breast cancer

Primary Breast Cancer Treatment Pathways

This study unearthed 30 primary breast cancer treatment pathways, as shown in Table 2. One-length models included surgery, radiotherapy, chemotherapy, and endocrine therapy. The surgical support was 0.8622658, indicating that 86.2% of patients had surgery. However, targeted therapy did not enter the frequent sequence mode, indicating that few patients use targeted therapy, and the proportion wasless than 15%. Of these sequence patterns, the longest had 4 items. According to the initial treatment, we divided the model of 2–4 items into neoadjuvant chemotherapy + surgery, surgery-based treatment, chemotherapy-based treatment, and other treatments (radiotherapy and endocrine treatment).

Table 2
Frequent sequence patterns of mining

Sequential mode	Support
chemotherapy	0.8827576
surgery	0.8622658
endocrine therapy	0.6756440
radiotherapy	0.3897834
neoadjuvant chemotherapy + surgery	
chemotherapy,surgery	0.1794496
surgery as the main treatment	
surgery,chemotherapy	0.7492681
surgery,endocrine therapy	0.5769906
surgery,chemotherapy,endocrine therapy	0.4972190
surgery,radiotherapy	0.3359192
surgery,chemotherapy,radiotherapy	0.3179157
surgery,radiotherapy,endocrine therapy	0.2442916
surgery,chemotherapy,radiotherapy,endocrine therapy	0.2312646
surgery,endocrine therapy,endocrine therapy	0.1847190
surgery,chemotherapy,endocrine therapy,endocrine therapy	0.1681792
surgery,endocrine therapy,radiotherapy	0.1656909
surgery,chemotherapy,endocrine therapy,radiotherapy	0.1538349
surgery,endocrine therapy,radiotherapy,endocrine therapy	0.1536885
chemotherapy as the main treatment	
chemotherapy,endocrine therapy	0.5682084
chemotherapy,radiotherapy	0.3628513
chemotherapy,chemotherapy	0.2697600
chemotherapy,endocrine therapy,endocrine therapy	0.2609778
chemotherapy,endocrine therapy,endocrine therapy	0.1964286
chemotherapy,endocrine therapy,radiotherapy	0.1768150
chemotherapy,chemotherapy,endocrine therapy	0.1680328
chemotherapy,endocrine therapy,radiotherapy,endocrine therapy	0.1624707
chemotherapy,chemotherapy,radiotherapy	0.1569087
other treatments (radiotherapy and endocrine therapy)	
radiotherapy,endocrine therapy	0.2808841
endocrine therapy,endocrine therapy	0.2565867

Sequential mode	Support
endocrine therapy,radiotherapy	0.2037471
endocrine therapy,radiotherapy,endocrine therapy	0.1826698

Table 2 shows that taking surgery as a reference, chronologically, chemotherapy can appear in patients before surgery (neoadjuvant chemotherapy), after surgery, or without surgery. Radiotherapy and endocrine therapy appeared postoperatively. Therefore, we identified the treatment stage of early breast cancer: preoperative stage (neoadjuvant chemotherapy) → surgery → postoperative chemotherapy → radiotherapy → endocrine therapy.

<Table 2 about here>

Table 2 Frequent sequence patterns of mining

Results of early breast cancer chemotherapy regimen clustering

Table 3 shows the results of clustering. Preoperative chemotherapy resulted in 3 types of chemotherapy, postoperative chemotherapy formed 3 types of chemotherapy, and nonoperative chemotherapy formed 2 types of chemotherapy.

Table 3
Clusters of chemotherapy drugs at different stages

drug	preoperative neoadjuvant chemotherapy	postoperative chemotherapy	chemotherapy in non-surgical patients
chemotherapy regimen1	Epirubicin Hydrochloride for Injection	Epirubicin Hydrochloride for Injection	Cyclophosphamide for injection
	Paclitaxel injection	Cyclophosphamide for injection	Epirubicin Hydrochloride for Injection
	Docetaxel injection	Docetaxel injection	Docetaxel injection
	Cyclophosphamide for injection	Fluorouracil injection(FLMD)	FluorouracilinjectionFLMD
	Fluorouracil injection(FLMD)	Fluorouracil injection(FNMD)	Pirarubicin hydrochloride for injection
	Carboplatin for injection	Paclitaxel injection	Capecitabine
		Carboplatin injection	Vinorelbine Bitartrate Injection
			Carmofluor tablets
			Tigio Capsules
			Gemcitabine hydrochloride for injection
			Vinorelbine Tartrate Injection
			Cisplatin injection
			Paclitaxel liposomes for injection
			Paclitaxel injection
			Carboplatin for injection
chemotherapy regimen2	Paclitaxel liposomes for injection	Gemcitabine hydrochloride for injection	Compound cyclophosphamide tablets
		Vinorelbine Tartrate Injection	Methotrexate for injection
		Loplatin for injection	Fluorouracil injection(FNMD)
		Cisplatin injection	
		Capecitabine	
		Vinorelbine Bitartrate Injection	
		Compound cyclophosphamide tablets	
		Methotrexate for injection	
chemotherapy regimen3	Gemcitabine hydrochloride for injection	Pirarubicin hydrochloride for injection	
	Vinorelbine Tartrate Injection	Ifosfamide for injection	
	Cisplatin injection	Vincristine sulfate for injection	

drug	preoperative neoadjuvant chemotherapy	postoperative chemotherapy	chemotherapy in non-surgical patients
	Capecitabine	Paclitaxel liposomes for injection	
		Doxorubicin Hydrochloride for Injection	

<Table 3 about here>

Table 3 Clusters of chemotherapy drugs at different stages

Secondary treatment pathways for early breast cancer

As a result, we found 55 rules for breast cancer treatment, which we use as secondary pathways for early breast cancer treatment. Table 4 shows the most supported sequence rules in different treatments and different numbers of items. From the number of items, the rules were between 2–5 items. According to the initial treatment, we included these rules into neoadjuvant chemotherapy + surgery, mainly surgical treatment, and radiotherapy + endocrine therapy. In each major category, they were sorted in descending order according to the number of items and the support for each rule. Neoadjuvant chemotherapy was mainly preoperative chemotherapy1. Surgical treatments after neoadjuvant chemotherapy were mainly radical surgery and modified radical surgery. Among the pathways with surgery as the main treatment, the ratio of modified radical surgery to postoperative chemotherapy 1 was 60.9%, which represented the highest proportion, followed by breast conservation surgery, radical surgery, and modified radical surgery + breast reconstruction surgery. Among the pathways dominated by radiotherapy and endocrine therapy, selective estrogen receptor modulators were the most commonly used endocrine regimens, followed by aromatase inhibitors.

Table 4
Mining sequence rules for breast cancer treatment

	items	sequence rule	Support	Confidence	Lift
		neoadjuvant chemotherapy + surgery			
	5-item	preoperative chemotherapy1,improved radical surgery□postoperative chemotherapy1□radiotherapy->selective estrogen receptor modulator	0.03193203	0.5000000	1.0334544
	4-item	preoperative chemotherapy1,selective estrogen receptor modulator□radiotherapy->selective estrogen receptor modulator	0.02856306	0.8590308	1.7755385
	3-item	preoperative chemotherapy1,postoperative chemotherapy1->radiotherapy	0.09140179	0.7222222	1.8515250
	2-item	preoperative chemotherapy1->improved radical surgery	0.10824667	0.7119461	0.9891037
		surgery as the main treatment			
Radical surgery	3-item	radical surgery,postoperative chemotherapy1->radiotherapy	0.03032079	0.6550633	1.6793530
	2-item	radical surgery->postoperative chemotherapy1	0.04628680	0.8777778	1.2381382
Improved radical surgery	4-item	improved radical surgery,selective estrogen receptor modulator□radiotherapy->selective estrogen receptor modulator	0.06444998	0.8644401	1.7867189
	3-item	improved radical surgery,supraclavicular lymph node dissection->postoperative chemotherapy1	0.06650066	0.8697318	1.2267891
	2-item	improved radical surgery->postoperative chemotherapy1	0.60524388	0.8408628	1.1860683
Breast-conserving surgery	4-item	breast-conserving surgery,selective estrogen receptor modulator□radiotherapy->selective estrogen receptor modulator	0.02314340	0.9132948	1.8876971
	3-item	breast-conserving surgery,postoperative chemotherapy1->radiotherapy	0.04452908	0.6846847	1.7552919
	2-item	breast-conserving surgery->postoperative chemotherapy1	0.06503589	0.6444122	0.9089674
Breast reconstruction	3-item	improved radical surgery,breast reconstruction->selective estrogen receptor modulator	0.02270397	0.5827068	1.2044018
	2-item	breast reconstruction->postoperative chemotherapy1	0.02768420	0.6847826	0.9659113
Other surgery	3-item	supraclavicular lymph node dissection,postoperative chemotherapy1->radiotherapy	0.04628680	0.5939850	1.5227696
	2-item	sentinel lymph node biopsy->postoperative chemotherapy1	0.08100190	0.7513587	1.0598194
Other	4-item	postoperative chemotherapy1,selective estrogen receptor modulator□radiotherapy->selective estrogen receptor modulator	0.08481031	0.8706767	1.7996094

	items	sequence rule	Support	Confidence	Lift
	3-item	selective estrogen receptor modulator,postoperative chemotherapy1->selective estrogen receptor modulator	0.03779112	0.5341615	1.1040631
		Radiation therapy + endocrine therapy as the main treatment			
	5-item	aromatase inhibitor,radiotherapy□aromatase inhibitor□radiotherapy->aromatase inhibitor	0.02182511	0.9490446	3.0065556
	4-item	radiotherapy,selective estrogen receptor modulator□radiotherapy->selective estrogen receptor modulator	0.05932328	0.8862144	1.8317245
	3-item	selective estrogen receptor modulator,radiotherapy->selective estrogen receptor modulator	0.11190860	0.7983281	1.6500714

<Table 4 about here>

Table 4 Mining sequence rules for breast cancer treatment

Results Of Cross-validation

Through 5-fold cross-validation, Pearson and Spearman correlation tests were performed. Table 5 shows the results of cross-validation. At the significance level of $P < 0.05$, all correlation coefficients of support, confidence and promotion were greater than 0.89. Using the Kolmogorov-Smirnov test, we found no significant differences between the sequence distributions.

Table 5
Cross-validation of sequence rules for early breast cancer treatment pathways

Set	Number of rules	Kolmogorov-Smirnov (*P*)			Pearson correlation			Spearman correlation		
		Supp	Conf	Lift	Supp	Conf	Lift	Supp	Conf	Lift
1	53	0.725	0.913	1.000	0.995***	0.925***	0.992***	0.925***	0.895***	0.960***
2	51	0.967	0.557	0.723	0.999***	0.956***	0.988***	0.980***	0.940***	0.955***
3	59	0.801	0.920	0.920	0.999***	0.953***	0.965***	0.944***	0.960***	0.970***
4	59	0.650	0.499	0.257	0.998***	0.948***	0.998***	0.974***	0.921***	0.969***
5	58	0.999	0.916	0.999	0.998***	0.971***	0.989***	0.971***	0.951***	0.978***
Average	56	0.828	0.761	0.780	0.998	0.951	0.986	0.959	0.933	0.966

<Table 5 about here>

Table 5 Cross-validation of sequence rules for early breast cancer treatment pathways

Discussion

In past research, there have been studies using frequent itemsets and association rule mining to infer relationships between drugs, laboratory results, and problems [30–32]. However, these data mining techniques cannot capture temporal information. Identifying all frequent sequence patterns in a transaction database, especially in a large electronic medical record database, requires effective algorithms to handle large search spaces. Sequential pattern mining is a data mining technique for identifying patterns of ordered events in a database. This method can be used for pattern recognition and prediction. Wright AP et al. [21] used sequential pattern mining to evaluate whether the method can effectively identify the time relationship between diabetes drugs and accurately predict the next drug that may be prescribed for patients. Tang C et al. [22]examined the use of sequential pattern mining techniques in large clinical datasets regarding the treatment and drug use patterns of childhood pneumonia. Zhan C et al. [23] discovered the side effects of drugs by mining the prescription sequence.

However, the above studies have explored the use of drugs, and cancer treatment is often a comprehensive treatment, including surgery, chemotherapy, radiotherapy, etc., and each treatment method comprises a variety of drugs, technologies and other components. The multidimensional and multilevel data make analysis difficult. How to discover such useful information for clinical treatment and hospital management is the challenge we face. To perform data mining at different granularities without generating too much information, we designed a three-step data mining method combining the cSPADE algorithm with cluster analysis to complete the mining of early breast cancer treatment paths. Through first-step sequence pattern mining, we found 30 frequent sequence patterns and determined the treatment stage of early breast cancer. Based on the determined treatment stage, we used reference guides and literature, expert consultation and cluster analysis to classify treatment options. Finally, we used the algorithm of sequential pattern mining again to mine 55 sequence rules as a secondary treatment path for early breast cancer. When clustering chemotherapeutic drugs, we considered previous studies' use of cluster analysis to obtain typical treatment plans. For example, Jingfeng C et al. [33] used AP clustering to extract typical treatment plans from EMR.

Through this study, we prove the effectiveness of sequential pattern mining, which can be used to determine the order of use of different treatments, to mine treatment paths from the real world and reconstruct stepwise usage pattern similar to those recommended. After mining the sequence pattern, we explored the stability of the mining results of the cSPADE algorithm for sequence rule mining of early breast cancer treatment paths through 5-fold cross-validation. Table 5 shows the number of patterns found in the iteration and the results of the statistical tests performed on each row: Kolmogorov-Smirnov, Pearson, and Spearman correlation tests for support, confidence, and lift. All measures found that in each iteration, all patterns were similar at a significance level of $P < 0.05$. All Pearson and Spearman correlation coefficients are higher than 0.75 [19], and the smallest correlation coefficient is 0.89, which means that the metric values in the training and validation sets are similar. The *p*-values of the Kolmogorov-Smirnov tests for the similarity of the distributions are all above 0.7, so no significant difference is found in the distributions.

Conclusion

Large-scale electronic medical record data are available, providing us with a unique opportunity to discover patterns using data mining methods. We combined the cSPADE algorithm and system clustering and adopted a three-stage mining process to realize the multilevel, vertical treatment mode of breast cancer. We cross-validated the sequence rules for early breast cancer treatment pathways and confirmed the stability of the results. Our results show that this approach can be used to generate potentially interesting and clinically meaningful cancer treatment sequence patterns and to manage clinical pathways in real-world cancer treatment, thereby guiding clinical practice.

Abbreviations

EMR

Electronic medical records; EHR:Electronic health records

Declarations

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The study obtained the approval by the the Ethics Committee, West China School of Medicine/West China Hospital, Sichuan University(approval number 2017-255). All information is obtained from the electronic hospital record. Informed consent was waived by institutional review board for this retrospective review study.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

All the authors declare that they have no conflicts of interest.

Funding

This study was supported by China's National Development and Reform Commission Grant 2018gfgw001 to Wei Zhang.

Authors' contributions

Conception and design: Qing Yang, Ting Luo, Wei Zhang. Analysis and interpretation: Qing Yang, Xiaorong Zhong. Data collection: Ting Luo, Ping He, Hong Zheng. Writing the article: Qing Yang. Critical revision of the article: Ting Luo, Wei Zhang. Final approval of the article: Qing Yang, Wei Zhang. Overall responsibility: Qing Yang, Wei Zhang. All authors have read and approved the final version of the article. Financial acquisition: Wei Zhang.

Acknowledgements

Not applicable.

Authors' information

[1]Institute of Hospital Management, West China Hospital, Sichuan University, Chengdu, China. [2]Department of Head, Neck and Mammary Gland Oncology, Cancer Center, West China Hospital, Sichuan University, Chengdu, China. [3]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China. [4]Laboratory of Molecular Diagnosis of Cancer, Clinical Research Center for Breast, West China Hospital, Sichuan University, Chengdu, China.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424.

2. Yamauchi C, Sekiguchi K, Nishioka A, et al. The Japanese Breast Cancer Society Clinical Practice Guideline for radiation treatment of breast cancer, 2015 edition. Breast Cancer. 2016;23(3):378–90.
3. Lefeuvre D, Le Bihan-Benjamin C, Pauporté I, et al. French Medico-Administrative Data to Identify the Care Pathways of Women With Breast Cancer. Clini Breast Cancer. 2017;17(4):e191–7.
4. Miguel RTD, Silvestre MAA, Imperial MLS,et al. Appraisal of the methodological quality of clinical practice guidelines in the Philippines. Int J Health Plann Manage. 2019;34(4):e1723–35.
5. Dy SM, Garg P, Nyberg D, et al. Critical Pathway Effectiveness: Assessing the Impact of Patient, Hospital Care, and Pathway Characteristics Using Qualitative Comparative Analysis. Health Serv Res. 2005;40(2):499–516.
6. Lawal AK, Rotter T, Kinsman L, et al. What is a clinical pathway? Refinement of an operational definition to identify clinical pathway studies for a Cochrane systematic review. BMC Med. 2016;14(1):35.
7. Rotter T, Kinsman L, James E, et al. The quality of the evidence base for clinical pathway effectiveness: Room for improvement in the design of evaluation trials. BMC Med Res Methodol. 2012;18(12):80.
8. ABIDI S. A Knowledge-Modeling Approach to Integrate Multiple Clinical Practice Guidelines to Provide Evidence-Based Clinical Decision Support for Managing Comorbid Conditions. J Med Syst. 2017;41(12):193.
9. Sharma DK, Solbrig HR, Tao C, et al. Building a semantic web-based metadata repository for facilitating detailed clinical modeling in cancer genome studies. J Biomed Semantics. 2017;8(1):19.
10. Wang HQ, Zhou TS, Tian LL, et al. Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. J Biomed Inform. 2014;52:354–63.
11. Claeson M, Hallberg S, HolmstrÖM P, et al. Modelling the future: System dynamics in the cutaneous malignant melanoma care pathway. Acta Derm Venereol. 2016;96(2):181–5.
12. Andrews R, Wynn M, Vallmuur K, et al. Leveraging Data Quality to Better Prepare for Process Mining: An Approach Illustrated Through Analysing Road Trauma Pre-Hospital Retrieval and Transport Processes in Queensland. Int J Environ Res Public Health. 2019;16(7):1138.
13. Zhang Y, Padman R, Patel N. Paving the COWpath: Learning and visualizing clinical pathways from electronic health record data. J Biomed Inform. 2015;58:186–97.
14. Pomares-quimbaya A, kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from electronic health records: A systematic review. BMC Med Res Methodol. 2019;19(1):155.
15. Korach ZT, Yang J, Rossetti SC, et al. Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. Int J Med Inf. 2020;135:204053.
16. Chen JH, Alagappan M, Goldstein MK, et al. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. Int J Med Inform. 2017;102 (2017): 71 – 9.
17. Huaulmé A, Harada K, ForestierHuaulmé G, et al. Sequential surgical signatures in micro-suturing task. Int J CARS. 2018;13(9):1419–28.
18. Huang Z, Shyu ML, Tien JM, et al. Prediction of Uterine Contractions Using Knowledge-Assisted Sequential Pattern Analysis. IEEE Trans Biomed Eng. 2013;60(5):1290–7.
19. Pimus I, Peleg M, Schertz M. Sequence mining of comorbid neurodevelopmental disorders using the SPADE algorithm. Methods Inf Med. 2016;55(3):223–33.
20. Davazdahemami B, Delen D. Examining the effect of prescription sequence on developing adverse drug reactions: The case of renal failure in diabetic patients. Int J Med Inform. 2019;125:62–70.
21. Wright AP, Wright AT, McCoy AB, et al. The use of sequential pattern mining to predict next prescribed medications. J Biomed Inform. 2015;53:73–80.
22. Tang C, Sun H, Xiong Y, et al. Medication use for childhood pneumonia at a children's hospital in Shanghai, China: Analysis of pattern mining algorithms. JMIR Med Inform. 2019;7(1):e12577.

23. Zhan C, Roughead E, Liu L, et al. A data-driven method to detect adverse drug events from prescription data. J Biomed Inform. 2018;85:10–20.
24. Niyazmand T, Izadi I. Pattern mining in alarm flood sequences using a modified PrefixSpan algorithm. ISA Trans. 2019;90:287–93.
25. Zihayat M, Davoudi H, An A. Mining significant high utility gene regulation sequential patterns. BMC Syst Biol. 2017;11(S6):109.
26. Furniss SK, Burton MM, Grando A, et al. Integrating process mining and cognitive analysis to study EHR workflow. AMIA Ann Symp proc. 2016;10:580-9.
27. Guo S, Li X, Liu H, et al. Integrating temporal pattern mining in ischemic stroke prediction and treatment pathway discovery for atrial fibrillation. AMIA Jt Summits Transl Sci Proc. 2017:122 – 30.
28. Wang Y, Hou W, Wang F. Mining co-occurrence and sequence patterns from cancer diagnoses in New York State. PLoS One. 2018;13(4):e0194407.
29. Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Mach Learn. 2001;42(1–2):31–60.
30. Tang JY, Chuang LY, Hsi E, et al. Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the apriori algorithm. Biomed Res Int. 2013:359634. https://doi.org/10.1155/2013/359634.
31. Murti S, Patwari GR, Joshi S. Knowledge discovery from medical data: Extracting influencing factors of breast cancer recurrences through predictive Apriori algorithm. Int J Adv Res Comp Sci. 2011;2(6):286.
32. Li J, Adilmagambetov A, Jabbar MSM, et al. On discovering co-location patterns in datasets: a case study of pollutants and child cancers. GeoInformatica. 2016;20(4):651–92.
33. Chen J, Sun L, Guo C, et al. A data-driven framework of typical treatment process extraction and evaluation. J Biomed Inform. 2018;83:178–95.

Figures

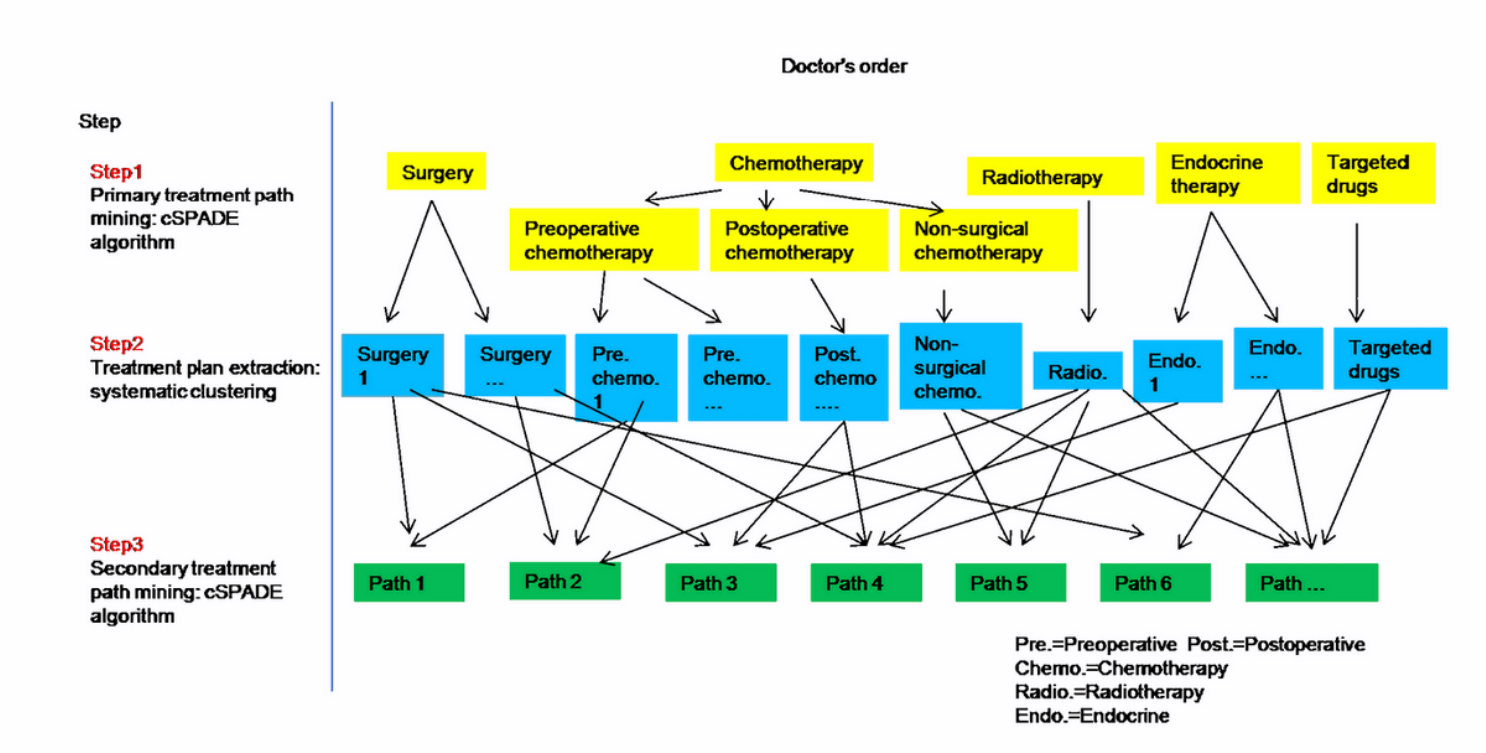

Figure 1

Mining process of early breast cancer treatment

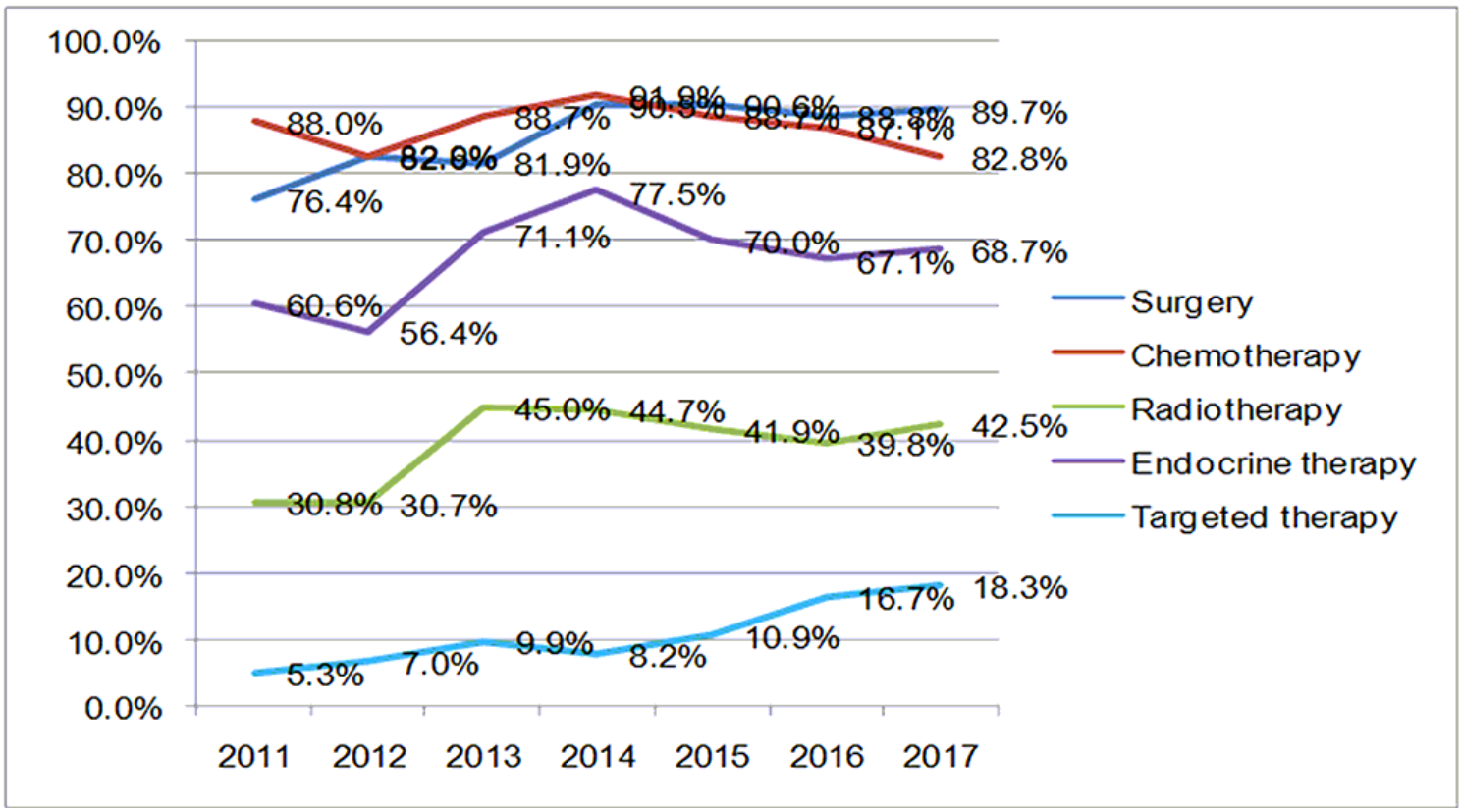

Figure 2

Trends in the use of different treatments for breast cancer