Association Rule Learning and Frequent Sequence Mining of Cancer Diagnoses in New York State

Yu Wang[1] and Fusheng Wang[1,2](✉)

[1] Department of Computer Science, Stony Brook University,
Stony Brook, NY 11794, USA
yuwang4@cs.stonybrook.edu, fusheng.wang@stonybrook.edu
[2] Department of Biomedical Informatics, Stony Brook University,
Stony Brook, NY 11794, USA

Abstract. Analyzing large scale diagnosis histories of patients could help to discover comorbidity or disease progression patterns. Recently, open data initiatives make it possible to access statewide patient data at individual level, such as New York State SPARCS data. The goal of this study is to explore frequent disease co-occurrence and sequence patterns of cancer patients in New York State using SPARCS data. Our collection includes 18,208,830 discharge records from 1,565,237 patients with cancer-related diagnoses during 2011–2015. We use Apriori algorithm to discover top disease co-occurrences for common cancer categories based on support. We generate top frequent sequences of diagnoses with at least one cancer related diagnosis from patients' diagnosis histories using the cSPADE algorithm. Our data driven approach provides essential knowledge to support the investigation of disease co-occurrence and progression patterns for improving the management of multiple diseases.

Keywords: Association rule learning · Sequence mining · SPARCS

1 Introduction

Disease co-occurrence, which means that two or more diseases co-occur within one patient [1], is a popular topic in public health studies. It sometimes represents comorbidity or multimorbidity and can suggest interactions between different risk factors like diagnoses, treatments and procedures [1,2]. Data mining and machine learning techniques are widely applied to public health domain to discover disease co-occurrences. For example, statistical methods can be used to measure the association between two different diagnoses [3], and structure learning models like Bayesian Network are used to analyze interactions in disease co-occurrence patterns [2]. Disease co-occurrences can also be identified by computing diseases that co-occur most frequently using Apriori-like algorithms [4]. Patterns and features discovered from comorbidities could provide a foundation

E. Begoli et al. (Eds.): DMAH 2017, LNCS 10494, pp. 121–135, 2017.
DOI: 10.1007/978-3-319-67186-4_10

for creating predictive models [5]. For example, comorbidities are informative features in predicting readmission risk of certain diseases [1].

Although disease co-occurrence is essential in studying correlations among different diseases, it fails to suggest temporal trends of diagnoses as information on the order in which diseases occur is not available. It therefore cannot reveal disease progression. Sequential data mining, which considers the order of data elements, has been used to detect temporal trends of various diseases. For example, windowing, episode rules and inductive logic programming are used to extract frequent sequential patterns of cardiovascular diseases [6]. Aggregate values and time intervals from health records are used as features to cluster patients into different cohorts [7]. Wavelet functions can help to analyze time series in healthcare data of patients with diabetes [8]. However, most of these methods are value-based, they use values from laboratory tests or other healthcare records to generate results. In our study, we adopt a sequence mining method that uses diagnosis codes (class labels) to study disease progression from patients' diagnosis histories.

Recently, open data initiatives from governments collect and make available large amounts of healthcare data, and provide a unique opportunity to study disease comorbidities and sequential patterns. They are attractive to researchers working on public health studies because of their completeness and inexpensive nature [9,10]. Such data are extensively used in healthcare research, such as prevention and detection of diseases, studying comorbidity and mortality, and advancing interventions, therapies and treatments [9]. They can also be combined with multiple data sources to serve different purposes, such as studying disease patterns and improving healthcare quality among different cohorts. For instance, predicting asthma-related emergency department visits [11] and analyzing temporal patterns of in-hospital falls among elderly patients [12].

As part of New York State's open data initiative, New York State Statewide Planning and Research Cooperative System (SPARCS) collects patient-level information on discharge records from hospitals, which contains patients' diagnosis, procedure and demographic information for over 35 years [13]. SPARCS is now widely applied to public health studies in New York State [14,15], such as correlations between various factors and outcomes of patients who suffer from different diseases [16–19], associations of different patient characteristics, diseases and treatments [16,20]. SPARCS is also used to discover temporal or spatial patterns of emergency department visits before, during and after Hurricane Sandy [21,22]. Researchers can benefit from SPARCS data by leveraging the long patient-level diagnosis histories, such as conducting population-based studies [23] and assessing completeness of disease reporting [24]. Patient-level longitudinal data can also embrace other data sources like drug exposure profiles and genetics data to study patterns in different cohorts [5].

The objective of this study is to find association rules (i.e., co-occurrences of diseases) and frequent sequence patterns from diagnosis histories of cancer patients in New York State using SPARCS data. Association rules learning of multiple diseases could imply comorbidities, while sequence patterns of diseases could indicate disease progression. We extract all discharge records of patients with at least one cancer-related diagnosis code, and convert the ninth and tenth revision of International Classification of Diseases (ICD-9 and ICD-10) diagnosis codes to single-level Clinical Classifications Software (CCS) diagnosis categories. The CCS cancer categories are used as disease labels in our work. We use Apriori algorithm for association rules learning to find potential comorbidities using multiple diagnoses from individual visits and cSPADE algorithm for frequent sequence mining to identify frequent disease sequence patterns from full discharge histories of patients in each cohort. We perform the studies by using only primary diagnoses and using all diagnoses (including secondary ones), to generate different patterns. We present the results based on several common cancer types, and we believe that the results will provide essential data and knowledge for clinical researchers to further investigate comorbidities and disease progression for improving the management of multiple diseases.

2 Methods

Using data mining and machine learning methods to study patients' profiles can help researchers to study comorbidities and disease progression [5]. Our objective is to conduct a patient-level longitudinal study using SPARCS data to discover frequent disease co-occurrence and sequence patterns. We first convert ICD-9

Table 1. Cancer-related CCS diagnosis categories and descriptions.

CCS	Description	CCS	Description
11	Cancer of head and neck	29	Cancer of prostate
12	Cancer of esophagus	30	Cancer of testis
13	Cancer of stomach	31	Cancer of other male genital organs
14	Cancer of colon	32	Cancer of bladder
15	Cancer of rectum and anus	33	Cancer of kidney and renal pelvis
16	Cancer of liver and intrahepatic bile duct	34	Cancer of other urinary organs
17	Cancer of pancreas	35	Cancer of brain and nervous system
18	Cancer of other GI organs; peritoneum	36	Cancer of thyroid
19	Cancer of bronchus; lung	37	Hodgkin‘s disease
20	Cancer; other respiratory and intrathoracic	38	Non-Hodgkin‘s lymphoma
21	Cancer of bone and connective tissue	39	Leukemias
22	Melanomas of skin	40	Multiple myeloma
23	Other non-epithelial cancer of skin	41	Cancer; other and unspecified primary
24	Cancer of breast	42	Secondary malignancies
25	Cancer of uterus	43	Malignant neoplasm without specification of site
26	Cancer of cervix	44	Neoplasms of unspecified nature or uncertain behavior
27	Cancer of ovary	45	Maintenance chemotherapy; radiotherapy
28	Cancer of other female genital organs		

and ICD-10 diagnosis codes to CCS diagnosis categories, and then use Apriori and cSPADE algorithms to identify patterns using these high-level categories. We only focus on histories of patients who have at least one of the cancer-related CCS diagnosis categories (Table 1).

2.1 Data Sources

We use SPARCS data and obtain histories of 21,466,868 patients from 97,849,071 discharge records during 2011–2015. Discharge records with all four kinds of claim types (i.e. inpatient, outpatient, ambulatory surgery and emergency department) are used to get a full history of each patient. Table 2 shows patient characteristics of our experiment data.

Table 2. Statistics of patient characteristics for selected cancer types.

Patient characteristics		Cancer						
		Lung and bronchus	Rectum and anus	Pancreas	Liver[a]	Non-Hodgkin's lymphoma	Prostate	Breast
Total number of patients		121,108	40,865	25,424	28,244	75,824	198,067	300,929
Age	<65	43,002	21,696	9,858	14,990	38,366	54,760	152,393
	65–74	38,120	9,492	7,333	7,304	17,513	64,154	69,886
	75–85	30,336	6,865	5,821	4,636	14,105	55,660	51,501
	>85	9,650	2,812	2,412	1,314	5,840	23,493	27,149
Sex	Male	58,320	20,818	12,694	17,348	39,102	198,067	3,919
	Female	62,785	20,043	12,729	10,896	36,719	0	297,004
	Unknown	3	4	1	0	3	0	6
Race	White	88,718	27,342	17,107	16,224	54,247	133,541	207,660
	Black or African American	12,490	5,022	3,369	3,979	6,948	31,120	34,221
	Native American or Alaskan Native	237	110	40	77	153	416	606
	Asian	3,713	1,428	800	2,062	1,647	3,000	8,483
	Native Hawaiian or Other Pacific Islander	224	64	35	50	101	375	532
	Other Race	13,934	6,212	3,710	5,446	11,358	26,512	43,372
	Unknown	1,792	687	363	406	1,370	3,103	6,055
Ethnicity	Spanish/Hispanic Origin	7,160	3,541	1,960	3,270	6,014	14,285	22,183
	Not of Spanish/Hispanic Origin	108,808	35,437	22,467	23,879	66,534	175,028	264,771
	Unknown	5,140	1,887	997	1,095	3,276	8,754	13,975

[a] Liver includes intrahepatic bile duct.

There are 25 data elements used to record ICD diagnosis codes of each hospital visit in SPARCS. The first diagnosis code is the primary diagnosis code that represents a main reason for a patient's hospital visit, the rest are secondary diagnosis codes that represent conditions coexist during that hospital visit. All ICD-9 and ICD-10 diagnosis codes are converted to their corresponding single-level CCS diagnosis categories, i.e. primary diagnosis categories and secondary

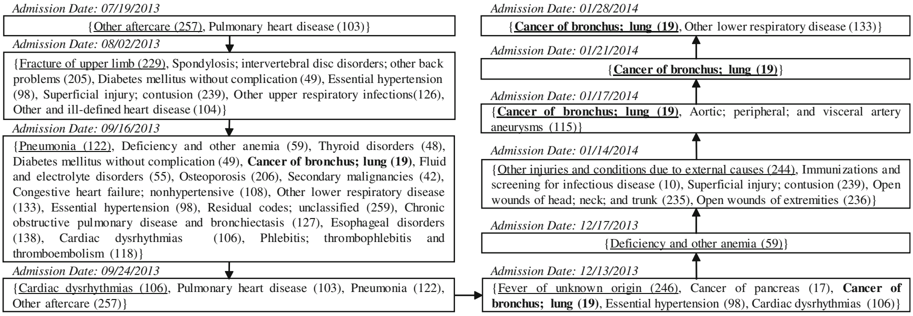
Admission Date: 07/19/2013
{Other aftercare (257), Pulmonary heart disease (103)}
Admission Date: 08/02/2013
{Fracture of upper limb (229), Spondylosis; intervertebral disc disorders; other back problems (205), Diabetes mellitus without complication (49), Essential hypertension (98), Superficial injury; contusion (239), Other upper respiratory infections(126), Other and ill-defined heart disease (104)}
Admission Date: 09/16/2013
{Pneumonia (122), Deficiency and other anemia (59), Thyroid disorders (48), Diabetes mellitus without complication (49), Cancer of bronchus; lung (19), Fluid and electrolyte disorders (55), Osteoporosis (206), Secondary malignancies (42), Congestive heart failure; nonhypertensive (108), Other lower respiratory disease (133), Essential hypertension (98), Residual codes; unclassified (259), Chronic obstructive pulmonary disease and bronchiectasis (127), Esophageal disorders (138), Cardiac dysrhythmias (106), Phlebitis; thrombophlebitis and thromboembolism (118)}
Admission Date: 09/24/2013
{Cardiac dysrhythmias (106), Pulmonary heart disease (103), Pneumonia (122), Other aftercare (257)}
Admission Date: 12/13/2013
{Fever of unknown origin (246), Cancer of pancreas (17), Cancer of bronchus; lung (19), Essential hypertension (98), Cardiac dysrhythmias (106)}
Admission Date: 12/17/2013
{Deficiency and other anemia (59)}
Admission Date: 01/14/2014
{Other injuries and conditions due to external causes (244), Immunizations and screening for infectious disease (10), Superficial injury; contusion (239), Open wounds of head; neck; and trunk (235), Open wounds of extremities (236)}
Admission Date: 01/17/2014
{Cancer of bronchus; lung (19), Aortic; peripheral; and visceral artery aneurysms (115)}
Admission Date: 01/21/2014
{Cancer of bronchus; lung (19)}
Admission Date: 01/28/2014
{Cancer of bronchus; lung (19), Other lower respiratory disease (133)}

Fig. 1. Diagnoses sequence of a patient with lung and bronchus cancer.

diagnosis categories. These high-level diagnosis categories are used to represent disease diagnoses to reduce dimensionality in data mining. We study patients with cancer diagnosis categories only. For each cancer category, patients whoever have at least one discharge record containing the cancer-related diagnosis information are selected into the cohort. There are 1,565,237 cancer patients and 18,208,830 history discharge records used in this study. Each patient's discharge records are grouped together using an encrypted unique patient identifier in SPARCS. Due to the length limit of this paper, we select seven types of cancers with high incident rates, which are consistent with the statistics by American Cancer Society [25], to present our results.

For each patient, discharge records are ordered by admission dates such that all CCS diagnosis categories on the same admission date form an element, and all elements are ordered to constitute a sequence (Fig. 1). Discharge records contain AIDS/HIV or abortion diagnoses are deleted from our experiment data because the admission dates are redacted and we cannot decide their positions in a sequence. An example of diagnoses sequence of a patient in cohort with lung and bronchus cancer is shown in Fig. 1. CCS diagnosis category descriptions reported on the same admission date are listed in brackets and form an element. The corresponding CCS category labels are marked in the parentheses following the descriptions. Admission dates are marked on top of each corresponding element. The primary diagnosis category of each element is underlined. CCS category that represents the targeted cancer (i.e., lung and bronchus cancer) is highlighted in bold.

2.2 Apriori Algorithm: Identifying Disease Co-occurrence Patterns

Association rule learning is a rule-based machine learning approach and is usually used to identify co-occurrences or temporal patterns between diseases in clinical domain [4]. In this study, we adopt Apriori algorithm [26] to identify disease co-occurrence patterns among each cohort. Only elements with targeted cancer CCS diagnosis categories are selected, and both primary and secondary diagnosis categories are used in our experiment. For instance, for the sequence illustrated in Fig. 1, elements where the targeted cancer CCS diagnosis categories are highlighted in bold are used.

Apriori algorithm discovers frequent disease co-occurrences by comparing their supports with a user-specified minimum support threshold. In Fig. 1, for example, if the support of pattern "{Cancer of bronchus; lung (19), Other lower respiratory disease (133)}" is 15%, it means that 15% of the elements in this cohort have this disease co-occurrence pattern. If the minimum support threshold is greater than 15%, this pattern will not be identified. However, if the minimum support threshold is set smaller than 15%, the pattern will be detected.

2.3 cSPADE Algorithm: Discovering Frequent Sequence Patterns

Because ICD diagnosis codes are the only data elements available in SPARCS that contain patient-level disease information, we can use frequent sequence mining [27] technique to find frequent disease sequence patterns among different cohorts. Since diagnosis codes are strictly ordered in sequences, the results might reveal disease progression. We use cSPADE algorithm [27] to discover frequent disease sequence patterns in different cohorts. We experiment on complete patient sequences with two settings: one is using only primary diagnosis categories, the other one is using both primary and secondary diagnosis categories. Figure 1 is an example of a complete patient sequence consists of both primary and secondary diagnosis categories. The length of a sequence pattern is the total number of elements in this sequence. There are 10 elements in the sequence in Fig. 1, thus it is a length-10 sequence.

cSPADE algorithm also works by comparing the support of a sequence pattern with the minimum support threshold. Multiple occurrences of a pattern in the same sequence is counted only once. For example, length-2 sequence pattern "{Cancer of bronchus; lung (19), Other lower respiratory disease (133)} → {Cardiac dysrhythmias (106)}" appears twice in Fig. 1, but this pattern will be counted only once in this sequence when calculating the support of this pattern. If the support of this sequence pattern is 15%, it means that the fraction of sequences containing this pattern in the targeted cohort is 15%. If the minimum support threshold is smaller than 15%, this sequence pattern is selected; otherwise the pattern is pruned in the searching results.

3 Results

We present the top five frequent disease co-occurrence and sequence patterns ranked by their supports in each cohort. Some meaningless results, such as patterns containing identical diagnosis categories, CCS diagnosis categories that represent unspecific disease groups or serve administrative purposes, patterns with length one and patterns irrelevant to targeted cancers, are filtered out when refining experiment results. We choose to present length-2 disease sequences in our experiment results, because longer disease sequence patterns obtained in our experiments usually contain repeated diagnosis categories that represent follow-up visits rather than disease progression.

Frequent disease co-occurrence patterns are presented in Table 3, and the results are generated using both primary and secondary diagnosis categories. Table 4 presents frequent disease sequence patterns discovered using only primary diagnosis categories. Table 5 demonstrates frequent disease sequence patterns identified using both primary and secondary diagnosis categories.

4 Discussion

4.1 Common CCS Categories in Different Cohorts

We can learn from Tables 3 and 5 that essential hypertension is the most frequent CCS diagnosis category among all results of either frequent disease co-occurrence or sequence patterns. However, essential hypertension appears in only three sequences in Table 4. This might because of the difference between primary diagnosis codes and secondary diagnosis codes in SPARCS data. Results in Tables 3 and 5 are generated using both primary and secondary diagnosis categories, but patterns in Table 4 are discovered using primary diagnosis categories only. Since primary diagnosis codes usually represent one major reason for a hospital visit and secondary diagnosis codes imply conditions that coexist during this visit, a combination of primary and secondary diagnosis codes usually contain richer diagnosis information. Perhaps cancers are more likely to be diagnosed with in the elderly and essential hypertension tend to be popular among old people, thus patients with cancer diagnoses could usually have essential hypertension. Combining primary and secondary diagnosis codes can help us easily detect this pattern. Disorders of lipid metabolism is another diagnosis category that is frequent in both Tables 3 and 5, while unseen in Table 4. The underlying theory might be similar.

4.2 Disparities Between Primary and Secondary Diagnosis Codes

Tables 4 and 5 both present frequent disease sequence patterns among different cohorts, while Table 4 shows the results produced using primary diagnosis categories only and Table 5 demonstrates results using both primary and secondary diagnosis categories. Frequent disease sequence patterns among same cohorts in these two tables are quite different. Disparities between Tables 4 and 5 could imply that either primary diagnosis codes or secondary diagnosis codes may be or may not be useful in finding potentially meaningful disease sequence patterns. Since primary diagnosis codes usually represent the main reason of a hospital visit, these codes are supposed to be good indicators of a patient's condition at admission. However, secondary diagnosis codes simply represent conditions that coexist in the same hospital visit, they might not be able to accurately represent a patient's condition responsible for that hospital visit. Thus, secondary diagnosis codes could be less meaningful information in this study. This can be justified by comparing results in Tables 4 and 5.

Table 3. Frequent disease co-occurrences for selected cancers, using both primary and secondary diagnosis categories.

	Lung and bronchus cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.2496
2	Screening and history of mental health and substance abuse codes	0.2271
3	Chronic obstructive pulmonary disease and bronchiectasis	0.1948
4	Disorders of lipid metabolism	0.1595
5	Coronary atherosclerosis and other heart disease	0.1136
	Rectum and anus cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.1908
2	Disorders of lipid metabolism	0.1088
3	Screening and history of mental health and substance abuse codes	0.0994
4	Deficiency and other anemia	0.0896
5	Diabetes mellitus without complication	0.0796
	Pancreas cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.2216
2	Diabetes mellitus without complication	0.1456
3	Fluid and electrolyte disorders	0.1260
4	Disorders of lipid metabolism	0.1234
5	Deficiency and other anemia	0.1065
	Liver and intrahepatic bile duct cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.2435
2	Hepatitis	0.2275
3	Diabetes mellitus without complication	0.1494
4	Screening and history of mental health and substance abuse codes	0.1340
5	Fluid and electrolyte disorders	0.1247
	Non-Hodgkin's lymphoma	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.1874
2	Deficiency and other anemia	0.1224
3	Disorders of lipid metabolism	0.1213
4	Screening and history of mental health and substance abuse codes	0.0910
5	Diabetes mellitus without complication	0.0835
	Prostate cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.3292
2	Disorders of lipid metabolism	0.2384
3	Coronary atherosclerosis and other heart disease	0.1739
4	Disorders of lipid metabolism, Essential hypertension	0.1521
5	Screening and history of mental health and substance abuse codes	0.1392
	Breast cancer	
	Top five frequent disease co-occurrences	Support
1	Essential hypertension	0.2159
2	Disorders of lipid metabolism	0.1302
3	Disorders of lipid metabolism, Essential hypertension	0.0851
4	Diabetes mellitus without complication	0.0829
5	Screening and history of mental health and substance abuse codes	0.0792

Table 4. Frequent sequence patterns for selected cancers, using primary diagnosis categories only.

	Lung and bronchus cancer	
	Top five frequent sequence patterns	Support
1	{Chronic obstructive pulmonary disease and bronchiectasis}→{Lung and bronchus cancer}	0.0700
2	{Pneumonia}→{Lung and bronchus cancer}	0.0578
3	{Lung and bronchus cancer}→{Pneumonia}	0.0567
4	{Lung and bronchus cancer}→{Chronic obstructive pulmonary disease and bronchiectasis}	0.0524
5	{Lung and bronchus cancer}→{Septicemia}	0.0520
	Rectum and anus cancer	
	Top five frequent sequence patterns	Support
1	{Rectum and anus cancer}→{Colon cancer}	0.1323
2	{Colon cancer}→{Rectum and anus cancer}	0.1206
3	{Rectum and anus cancer}→{Complications of surgical procedures or medical care}	0.0521
4	{Gastrointestinal hemorrhage}→{Rectum and anus cancer}	0.0509
5	{Abdominal pain}→{Rectum and anus cancer}	0.0479
	Pancreas cancer	
	Top five frequent sequence patterns	Support
1	{Pancreatic disorders}→{Pancreas cancer}	0.1256
2	{Abdominal pain}→{Pancreas cancer}	0.0994
3	{Biliary tract disease}→{Pancreas cancer}	0.0914
4	{Pancreas cancer}→{Septicemia}	0.0794
5	{Pancreas cancer}→{Abdominal pain}	0.0618
	Liver and intrahepatic bile duct cancer	
	Top five frequent sequence patterns	Support
1	{Hepatitis}→{Liver and intrahepatic bile duct cancer}	0.0711
2	{Abdominal pain}→{Liver and intrahepatic bile duct cancer}	0.0688
3	{Liver and intrahepatic bile duct cancer}→{Hepatitis}	0.0586
4	{Liver and intrahepatic bile duct cancer}→{Septicemia (except in labor)}	0.0528
5	{Biliary tract disease}→{Liver and intrahepatic bile duct cancer}	0.0458
	Non-Hodgkin's lymphoma	
	Top five frequent sequence patterns	Support
1	{Lymphadenitis}→{Non-Hodgkin's lymphoma}	0.0458
2	{Non-Hodgkin's lymphoma}→{Septicemia (except in labor)}	0.0458
3	{Non-Hodgkin's lymphoma}→{Deficiency and other anemia}	0.0433
4	{Deficiency and other anemia}→{Non-Hodgkin's lymphoma}	0.0415
5	{Non-Hodgkin's lymphoma}→{Diseases of white blood cells}	0.0368
	Prostate cancer	
	Top five frequent sequence patterns	Support
1	{Prostate cancer}→{Genitourinary symptoms and ill-defined conditions}	0.0424
2	{Genitourinary symptoms and ill-defined conditions}→{Prostate cancer}	0.0401
3	{Essential hypertension}→{Prostate cancer}	0.0323
4	{Prostate cancer}→{Essential hypertension}	0.0292
5	{Spondylosis; intervertebral disc disorders; other back problems}→{Prostate cancer}	0.0283
	Breast cancer	
	Top five frequent sequence patterns	Support
1	{Nonmalignant breast conditions}→{Breast cancer}	0.0965
2	{Breast cancer}→{Nonmalignant breast conditions}	0.0796
3	{Spondylosis; intervertebral disc disorders; other back problems}→{Breast cancer}	0.0353
4	{Breast cancer}→{Spondylosis; intervertebral disc disorders; other back problems}	0.0331
5	{Essential hypertension}→{Breast cancer}	0.0295

Table 5. Frequent sequence patterns for selected cancers, using both primary and secondary diagnosis categories.

Lung and bronchus cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Lung and bronchus cancer}	0.5377
2	{Screening and history of mental health and substance abuse codes}→{Lung and bronchus cancer}	0.5240
3	{Lung and bronchus cancer}→{Essential hypertension}	0.4508
4	{Lung and bronchus cancer}→{Screening and history of mental health and substance abuse codes}	0.4350
5	{Disorders of lipid metabolism}→{Lung and bronchus cancer}	0.4044

Rectum and anus cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Rectum and anus cancer}	0.4338
2	{Rectum and anus cancer}→{Colon cancer}	0.4273
3	{Rectum and anus cancer}→{Essential hypertension}	0.4218
4	{Colon cancer}→{Rectum and anus cancer}	0.3863
5	{Disorders of lipid metabolism}→{Rectum and anus cancer}	0.2931

Pancreas cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Pancreas cancer}	0.5440
2	{Pancreas cancer}→{Essential hypertension}	0.4156
3	{Disorders of lipid metabolism}→{Pancreas cancer}	0.3857
4	{Essential hypertension}→{Essential hypertension, Pancreas cancer}	0.3767
5	{Fluid and electrolyte disorders}→{Pancreas cancer}	0.3718

Liver and intrahepatic bile duct cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Liver and intrahepatic bile duct cancer}	0.5116
2	{Liver and intrahepatic bile duct cancer}→{Essential hypertension}	0.4205
3	{Liver and intrahepatic bile duct cancer}→{Fluid and electrolyte disorders}	0.3510
4	{Screening and history of mental health and substance abuse codes}→{Liver and intrahepatic bile duct cancer}	0.3439
5	{Fluid and electrolyte disorders}→{Liver and intrahepatic bile duct cancer}	0.3271

Non-Hodgkin's lymphoma

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Non-Hodgkin's lymphoma}	0.4237
2	{Non-Hodgkin's lymphoma}→{Essential hypertension}	0.3911
3	{Disorders of lipid metabolism}→{Non-Hodgkin's lymphoma}	0.3128
4	{Deficiency and other anemia}→{Non-Hodgkin's lymphoma}	0.2925
5	{Non-Hodgkin's lymphoma}→{Deficiency and other anemia}	0.2920

Prostate cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Prostate cancer}	0.4849
2	{Prostate cancer}→{Essential hypertension}	0.4667
3	{Disorders of lipid metabolism}→{Prostate cancer}	0.3707
4	{Prostate cancer}→{Disorders of lipid metabolism}	0.3654
5	{Genitourinary symptoms and ill-defined conditions}→{Prostate cancer}	0.2156

Breast cancer

	Top five frequent sequence patterns	Support
1	{Essential hypertension}→{Breast cancer}	0.3958
2	{Breast cancer}→{Essential hypertension}	0.3887
3	{Disorders of lipid metabolism}→{Breast cancer}	0.2804
4	{Breast cancer}→{Disorders of lipid metabolism}	0.2779
5	{Nonmalignant breast conditions}→{Breast cancer}	0.2177

For patients with lung and bronchus cancer in Table 4, the most frequent sequences mainly consist of respiratory system diseases, such as pneumonia and chronic obstructive pulmonary disease and bronchiectasis. But there is no respiratory system disease in the top five frequent disease sequence patterns among the same patient cohort in Table 5. Another typical cohort is patients with liver and intrahepatic bile duct cancer. We can learn from Table 4 that patients in this cohort sometimes expose themselves to hepatitis or biliary tract disease. However, such patterns are not available in Table 5. Also for patients with Non-Hodgkin's lymphoma, frequent sequence patterns shown in Tables 4 and 5 are quite different. Only results in Table 4 capture the existence of lymphadenitis and disease of white blood cells. Also, the most frequent disease sequence patterns among this cohort in Table 4 all consist of immune system diseases.

4.3 Frequent Disease Co-occurrence Patterns Versus Frequent Disease Sequence Patterns

One major difference between disease sequence and co-occurrence patterns is that the orders of diagnoses are taken into consideration in a disease sequence, while disease co-occurrences simply represent different diagnoses that occur simultaneously. Disease sequence pattern can therefore be a potential indicator of disease progression. Since the order of two different diagnosis categories is the major factor to consider when tracking disease progression, we retain a frequent disease sequence pattern in the results, if its elements are reversed in another top frequent disease sequence pattern.

For instance, sequence patterns "{Rectum and anus cancer} → {Colon cancer}" and "{Colon cancer} → {Rectum and anus cancer}" are both kept in Table 4. The former has support 0.1323, which is slightly greater than the latter (0.1206). Perhaps it is because that rectum and anus cancer are more likely to develop into colon cancer, but fewer patients suffer from colon cancer can eventually have rectum and anus cancer. There could be causal relationships between the two diseases, or perhaps it is simply a result of the different mechanisms of these two types of cancers.

Another typical pattern is in disease sequences containing essential hypertension. In Table 5, for example, sequence pattern "{Essential hypertension} → {Pancreas cancer}" has support 0.5440, which is higher than the reversed sequence "{Pancreas cancer} → {Essential hypertension}" with support 0.4156. It is evident that all the sequences where essential hypertension is at the first position have higher supports than their reversed sequences. It is an interesting phenomenon that perhaps imply the progression of pancreas cancer. However, we cannot obtain any information on disease progression from disease co-occurrence patterns. For example, Table 3 shows that pattern "{Pancreas cancer, Essential hypertension}" is with the highest support among patients with pancreas cancer. It simply suggests that these two diagnoses co-occur frequently, but no information on the order in which they occur is available.

4.4 Validation of Results

Many public health studies use data from only one or a few hospitals collected in a short period of time [3,4,10]. However, SPARCS has been collecting more representative and comprehensive data for over 35 years, as all Article 28 facilities (i.e. hospitals, nursing homes, and diagnostic treatment centers) certified for inpatient care and all facilities providing ambulatory surgery services in New York State are required to submit inpatient or outpatient data to SPARCS [13]. We therefore have a large-scale dataset with longer patient histories that could help generate potentially meaningful results.

For disease co-occurrences (Table 3), patients with lung and bronchus cancer usually have chronic obstructive pulmonary disease and bronchiectasis observed at the same time. Since these two diseases are both respiratory system diseases, they are reasonably correlated with each other. The same applies for patients with pancreas cancer. Patients in this cohort have a risk of suffering from diabetes, as pancreas cancer and diabetes are clinically correlated [28]. Moreover, patients with liver and intrahepatic bile duct cancer also have chance to be diagnosed with hepatitis at the same time, because these two diseases are also associated with each other [28].

As for disease sequences (Table 4), many patients have pancreatic disorders (not diabetes) or biliary tract disease before being diagnosed with pancreas cancer. This might be a typical disease progression pattern in clinical studies and could help domain experts to identify pancreas cancer in the early stages. Another representative result is about Non-Hodgkin's lymphoma, because the result sequences usually consist of immune system diseases. The top sequence patterns suggest that lymphadenitis is likely to happen before Non-Hodgkin's lymphoma and disease of white blood cells is usually diagnosed after Non-Hodgkin's lymphoma.

Although secondary diagnosis codes could be redundant information on patient conditions, they are also able to produce some potentially interesting and meaningful patterns on disease progression when combined with primary diagnosis codes. For example, prostate cancer is more likely to be diagnosed after genitourinary symptoms and ill-defined conditions are identified, and breast cancer usually happens after nonmalignant breast conditions (Table 5). Since these two patterns have comparatively higher supports than other sequence patterns in the same cohorts, they could be typical patterns in clinical studies.

5 Conclusion

We employ association rule learning (Apriori algorithm) and frequent sequence mining (cSPADE algorithm) to identify frequent disease co-occurrence and sequence patterns among cancer patients using SPARCS data. Different types of diagnosis codes are utilized in our experiments. Seven cohorts where cancers are with high incident rates are selected to present the results. Our results suggest that the methods adopted can generate potentially interesting and clinically meaningful disease co-occurrence and sequence patterns. These patterns might be able to imply comorbidities and disease progression. However, due to the limitation of information that diagnosis codes can convey in SPARCS, our results contain some redundant or less meaningful patterns irrelevant to the targeted cancers. Since SPARCS is designed to serve administrative purpose to monitor and improve qualities of hospital services and data reporting, we believe our study could not only help to improve healthcare qualities provided to serve cancer patients, but also throw light upon researches using diagnosis codes in SPARCS.

Since high-level diagnosis categories contain richer but less specific diagnoses information than diagnosis codes, we can use low-level ICD-9 and ICD-10 diagnosis codes in our future researches to see if more specific and useful patterns can be extracted. We can also experiment on a cohort with one certain disease to narrow down the scope of our study and gain a deeper insight into that specific cohort.

Acknowledgments. This work is supported in part by NSF ACI 1443054, by NSF IIS 1350885 and by NSF IIP1069147.

References

1. Stiglic, G., Brzan, P.P., Fijacko, N., Wang, F., Delibasic, B., Kalousis, A., Obradovic, Z.: Comprehensible predictive modeling using regularized logistic regression and comorbidity based features. PLoS ONE **10**(12), e0144439 (2015). doi:10.1371/journal.pone.0144439
2. Lappenschaar, M., Hommersom, A., Lagro, J., Lucas, P.J.: Understanding the co-occurrence of diseases using structure learning. In: Conference on Artificial Intelligence in Medicine in Europe, pp. 135–144 (2013). doi:10.1007/978-3-642-38326-7_21
3. Munson, M.E., Wrobel, J.S., Holmes, C.M., Hanauer, D.A.: Data mining for identifying novel associations and temporal relationships with Charcot foot. J. Diabetes Res. (2014). doi:10.1155/2014/214353
4. Kost, R., Littenberg, B., Chen, E.S.: Exploring generalized association rule mining for disease co-occurrences. In: AMIA Annual Symposium Proceedings 2012, p. 1284 (2012)
5. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. **13**(6), 395–405 (2012). doi:10.1038/nrg3208
6. Kléma, J., Nováková, L., Karel, F., Stepankova, O., Zelezny, F.: Sequential data mining: a comparative case study in development of atherosclerosis risk factors. IEEE Trans. Syst. Man Cybern. Part C (Applications and Reviews) **38**(1), 3–15 (2008). doi:10.1109/tsmcc.2007.906055
7. Baxter, R.A., Williams, G.J., He, H.: Feature selection for temporal health records. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 198–209 (2001). doi:10.1007/3-540-45357-1_24
8. Lin, W., Orgun, M.A., Williams, G.J.: Mining temporal patterns from health care data. In: International Conference on Data Warehousing and Knowledge Discovery, pp. 222–231 (2002). doi:10.1007/3-540-46145-0_22

9. Ferver, K., Burton, B., Jesilow, P.: The use of claims data in healthcare research. Open Public Health J. **2**, 11–24 (2009). doi:10.2174/1874944500902010011
10. Tyree, P.T., Lind, B.K., Lafferty, W.E.: Challenges of using medical insurance claims data for utilization analysis. Am. J. Med. Qual. **21**(4), 269–275 (2006). doi:10.1177/1062860606288774
11. Ram, S., Zhang, W., Williams, M., Pengetnze, Y.: Predicting asthma-related emergency department visits using big data. IEEE J. Biomed. Health Inform. **19**(4), 1216–1223 (2015). doi:10.1109/jbhi.2015.2404829
12. López-Soto, P.J., Smolensky, M.H., Sackett-Lundeen, L.L., De Giorgi, A., Rodríguez-Borrego, M.A., Manfredini, R., Pelati, C., Fabbian, F.: Temporal patterns of in-hospital falls of elderly patients. Nurs. Res. **65**(6), pp. 435–445 (2016). doi:10.1097/nnr.0000000000000184
13. Statewide Planning and Research Cooperative System (SPARCS). https://www.health.ny.gov/statistics/sparcs/
14. Chen, X., Wang, F.: Integrative spatial data analytics for public health studies of new york state. In: AMIA Annual Symposium Proceedings, vol. 2016, p. 391 (2016)
15. Chen, X., Wang, Y., Schoenfeld, E., Saltz, M., Saltz, J., Wang, F.: Spatio-temporal analysis for New York State SPARCS data. In: Proceedings of 2017 AMIA Joint Summits on Translational Science (2017)
16. Bekelis, K., Missios, S., Coy, S., Rahmani, R., Singer, R.J., MacKenzie, T.A.: Surgical clipping versus endovascular intervention for the treatment of subarachnoid hemorrhage patients in New York State. PLoS ONE **10**(9), e0137946 (2015). doi:10.1371/journal.pone.0137946
17. Missios, S., Bekelis, K.: Regional disparities in hospitalization charges for patients undergoing craniotomy for tumor resection in New York State: correlation with outcomes. J. Neurooncol. **128**(2), 365–371 (2016). doi:10.1007/s11060-016-2122-0
18. Bekelis, K., Missios, S., Coy, S., MacKenzie, T.A.: Scope of practice and outcomes of cerebrovascular procedures in children. Child's Nerv. Syst. **32**(11), 2159–2164 (2016). doi:10.1007/s00381-016-3114-2
19. Bekelis, K., Missios, S., Coy, S., MacKenzie, T.A.: Comparison of outcomes of patients with inpatient or outpatient onset ischemic stroke. J. Neurointerventional Surg., pp. neurintsurg-2015 (2016). doi:10.1136/neurintsurg-2015-012145
20. Dy, C.J., Lane, J.M., Pan, T.J., Parks, M.L., Lyman, S.: Racial and socioeconomic disparities in hip fracture care. J. Bone Joint Surg. Am. **98**(10), 858–865 (2016)
21. Kim, H., Schwartz, R.M., Hirsch, J., Silverman, R., Liu, B., Taioli, E.: Effect of Hurricane Sandy on Long Island emergency departments visits. Disaster Med. Public Health Preparedness **10**(03), 344–350 (2016). doi:10.1017/dmp.2015.189
22. He, F.T., De La Cruz, N.L., Olson, D., Lim, S., Seligson, A.L., Hall, G., Jessup, J., Gwynn, C.: Temporal and spatial patterns in utilization of mental health services during and after hurricane sandy: emergency department and inpatient hospitalizations in New York City. Disaster Med. Public Health Preparedness **10**(03), 512–517 (2016). doi:10.1017/dmp.2016.89
23. Hodgins, J.L., Vitale, M., Arons, R.R., Ahmad, C.S.: Epidemiology of medial ulnar collateral ligament reconstruction: a 10-year study in New York State. Am. J. Sports Med. **44**(3), 729–734 (2016). doi:10.1177/0363546515622407
24. Arakaki, L., Ngai, S., Weiss, D.: Completeness of Neisseria meningitidis reporting in New York City, 19892010. Epidemiol. Infect. **144**(11), 2374–2381 (2016). doi:10.1017/s0950268816000406
25. Cancer facts & figures 2017. American Cancer Society (2017)

26. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
27. Zaki, M.J.: Sequence mining in categorical domains: incorporating constraints. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, pp. 422–429 (2000). doi:10.1145/354756.354849
28. Mayo Clinic. http://www.mayoclinic.org