RESEARCH ARTICLE

Mining co-occurrence and sequence patterns from cancer diagnoses in New York State

Yu Wang[1], **Wei Hou**[2], **Fusheng Wang**[1,3]*

1 Department of Computer Science, Stony Brook University, Stony Brook, New York, United States of America, **2** Department of Family, Population and Preventive Medicine, Stony Brook University, Stony Brook, New York, United States of America, **3** Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, United States of America

* fusheng.wang@stonybrook.edu

OPEN ACCESS

Citation: Wang Y, Hou W, Wang F (2018) Mining co-occurrence and sequence patterns from cancer diagnoses in New York State. PLoS ONE 13(4): e0194407. https://doi.org/10.1371/journal.pone.0194407

Editor: Francesco Russo, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, DENMARK

Received: July 11, 2017

Accepted: February 19, 2018

Published: April 26, 2018

Copyright: © 2018 Wang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The dataset is stored with SPARCS as limited data that cannot be publicly shared because it contains indirect identifiers that are deemed identifiable according to Health Insurance Portability and Accountability Act (HIPAA) standards. Data access requests for limited data are reviewed by SPARCS Operations staff. Data access information can be found at: https://www.health.ny.gov/statistics/sparcs/access/. The contact information of SPARCS team is as follows: Questions/Comments: sparcs.

Abstract

The goal of this study is to discover disease co-occurrence and sequence patterns from large scale cancer diagnosis histories in New York State. In particular, we want to identify disparities among different patient groups. Our study will provide essential knowledge for clinical researchers to further investigate comorbidities and disease progression for improving the management of multiple diseases. We used inpatient discharge and outpatient visit records from the New York State Statewide Planning and Research Cooperative System (SPARCS) from 2011-2015. We grouped each patient's visit history to generate diagnosis sequences for seven most popular cancer types. We performed frequent disease co-occurrence mining using the Apriori algorithm, and frequent disease sequence patterns discovery using the cSPADE algorithm. Different types of cancer demonstrated distinct patterns. Disparities of both disease co-occurrence and sequence patterns were observed from patients within different age groups. There were also considerable disparities in disease co-occurrence patterns with respect to different claim types (i.e., inpatient, outpatient, emergency department and ambulatory surgery). Disparities regarding genders were mostly found where the cancer types were gender specific. Supports of most patterns were usually higher for males than for females. Compared with secondary diagnosis codes, primary diagnosis codes can convey more stable results. Two disease sequences consisting of the same diagnoses but in different orders were usually with different supports. Our results suggest that the methods adopted can generate potentially interesting and clinically meaningful disease co-occurrence and sequence patterns, and identify disparities among various patient groups. These patterns could imply comorbidities and disease progressions.

Introduction

Background and significance

Patient level longitudinal data mining and pattern discovery is a common approach for public health studies. For example, disease co-occurrence patterns and disease sequence patterns from large number of patients' diagnosis histories could help to discover comorbidity or

submissions@health.ny.gov; Data Requests: sparcs.requests@health.ny.gov.

Funding: This work was supported by National Science Foundation-Advanced Cyberinfrastructure (ACI)-1443054 (https://www.nsf.gov/cise/oac/about.jsp, received by FW); National Science Foundation-Information and Intelligent Systems (IIS)-1350885 (https://www.nsf.gov/div/index.jsp?div=IIS, received by FW); National Science Foundation-Industrial Innovation and Partnerships (IIP)-1069147 (https://www.nsf.gov/div/index.jsp?div=IIP), received by FW).

Competing interests: The authors have declared that no competing interests exist.

disease progression patterns. Generally, disease co-occurrence patterns could imply comorbidities and disease sequence patterns could reveal disease progression. For instance, novel associations or generalized associations from raw data helped explore co-occurrence patterns of multiple diseases [1, 2]. Event sequences were mapped to a general knowledge representation model for mining longitudinal event data [3]. Specific techniques like windowing, episode rules and inductive logic programming were applied to identify sequential or temporal patterns related to cardiovascular diseases [4]. Sequential patterns were also helpful in analyzing temporal trends of diseases in certain situations [5]. However, due to the limitation of data availability, most previous studies were based on small datasets or a limited number of healthcare facilities.

Recently, open data initiatives from governments make available large amounts of healthcare data and provide researchers with a unique opportunity to study disease co-occurrence and sequence patterns. For example, an early project worked on discovering disease progression and mitigating potential adverse outcomes of chronic obstructive pulmonary disease (COPD) [6]. Another work used machine learning techniques such as hidden Markov models to study disparities among different cohorts by clustering patients into different groups using patient claim data [7]. Healthcare data were also combined with multiple data sources, such as social economic data and social media data, to predict hospital visits regarding certain diseases in specific areas [8].

As a representative of New York State's open data initiative, the Statewide Planning and Research Cooperative System (SPARCS) [9] has been collecting patient level demographic and diagnosis information on discharges for over 35 years. All article 28 licensed facilities (i.e. hospitals, nursing homes, and diagnostic treatment centers) in New York State are required to report outpatient, inpatient, emergency department and ambulatory surgery discharge records to SPARCS every year [10]. It has already been used in big data analysis in medical domain, like evaluating the quality of data reporting [11] and discovering associations between various factors and outcomes of treatments [12–14]. Besides, SPARCS provides rich data for exploring spatial, temporal and spatio-temporal patterns of various diagnoses [15–18].

Objective

This is a retrospective cohort study aiming at analyzing frequent disease co-occurrence and sequence patterns of cancer diagnoses in New York State. We studied disparities among these frequent patterns with respect to age, gender and claim types (i.e., inpatient, outpatient, emergency department and ambulatory surgery) for hospital visits or stays. Since cancer ranks the second in the leading causes of deaths in the United States [19], 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 cancers.

Materials and methods

This study has been approved by Stony Brook University IRB (CORIHS B). We took advantage of the cancer-related diagnosis information available in SPARCS data, i.e., the ninth and tenth revision of International Classification of Diseases (ICD-9 and ICD-10) diagnosis codes, and converted them to single-level Clinical Classifications Software (CCS) diagnosis categories [20] to discover disease co-occurrence and sequence patterns from patients' full diagnosis histories within a five-year time frame.

Data sources

While our SPARCS data from most claim types are available as early as the year 2003, outpatient records are only available since 2011. To provide a comprehensive history of patient visits, we choose discharge records in SPARCS during 2011-2015 where all claim types are available. Descriptive statistics of cancer patients based on discharge records in SPARCS during 2011-2015 are presented in Table 1. Each discharge record contains one or more ICD-9 or ICD-10 diagnosis codes. The first diagnosis code is the primary diagnosis code that represents the main reason for that hospital visit. The rest are secondary diagnosis codes that represent the conditions coexisting during the same hospital stay or visit. To reduce dimensionality of data, we mapped ICD diagnosis codes to single-level CCS diagnosis categories and used CCS

Table 1. Patient characteristics of seven types of cancer in SPARCS, 2011-2015.

Patient characteristics	Cancer						
	Lung and bronchus	Rectum and anus	Pancreas	Liver*	Non-Hodgkin's lymphoma	Prostate	Breast
Total population, n	120,833	40,816	25,352	28,190	75,718	197,847	300,682
Age, years							
Mean (std)	68.37 (12.16)	63.07 (14.70)	67.72 (13.14)	63.37 (13.89)	61.81 (18.05)	71.27 (10.90)	64.26 (14.49)
Median (min, max)	69 (0, 111)	63 (0, 102)	68 (0, 121)	64 (0, 106)	64 (0, 112)	71 (0, 111)	64 (0, 124)
<=34, n (%)	1,071 (0.89)	1,367 (3.35)	305 (1.20)	828 (2.94)	6,699 (8.85)	279 (0.14)	4,993 (1.66)
35-54, n (%)	14,397 (11.91)	10,146 (24.86)	3,537 (13.95)	5,251 (18.63)	15,681 (20.71)	12,148 (6.14)	76,410 (25.41)
55-74, n (%)	65,479 (54.19)	19,643 (48.13)	13,313 (52.51)	16,177 (57.39)	33,440 (44.16)	106,400 (53.78)	140,758 (46.81)
>=75, n (%)	39,886 (33.01)	9,660 (23.67)	8,197 (32.33)	5,934 (21.05)	19,898 (26.28)	79,020 (39.94)	78,521 (26.11)
Gender							
Male, n (%)	58,179 (48.15)	20,791 (50.94)	12,669 (49.97)	17,311 (61.41)	39,031 (51.55)	197,847 (100.00)	3,914 (1.30)
Female, n (%)	62,651 (51.85)	20,021 (49.05)	12,682 (50.02)	10,879 (38.59)	36,684 (48.45)	0 (0.00)	296,762 (98.70)
Race							
White, n (%)	88,497 (73.24)	27,316 (66.92)	17,054 (67.27)	16,201 (57.47)	54,159 (71.53)	133,385 (67.42)	207,472 (69.00)
Black or African American, n (%)	12,454 (10.31)	5,014 (12.28)	3,357 (13.24)	3,972 (14.09)	6,938 (9.16)	31,067 (15.70)	34,199 (11.37)
Native American or Alaskan Native, n (%)	236 (0.20)	110 (0.27)	40 (0.16)	78 (0.28)	153 (0.20)	415 (0.21)	607 (0.20)
Asian, n (%)	3,707 (3.07)	1,426 (3.49)	801 (3.16)	2,056 (7.29)	1,643 (2.17)	2,994 (1.51)	8,478 (2.82)
Native Hawaiian or Other Pacific Islander, n (%)	225 (0.19)	65 (0.16)	35 (0.14)	50 (0.18)	101 (0.13)	375 (0.19)	531 (0.18)
Other Race, n (%)	13,926 (11.52)	6,199 (15.19)	3,704 (14.61)	5,430 (19.26)	11,349 (14.99)	26,504 (13.40)	43,359 (14.42)
Ethnicity							
Spanish/Hispanic Origin, n (%)	7,154 (5.92)	3,533 (8.66)	1,956 (7.72)	3,257 (11.55)	6,002 (7.93)	14,267 (7.21)	22,173 (7.37)
Not of Spanish/Hispanic Origin, n (%)	108,549 (89.83)	35,403 (86.74)	22,403 (88.37)	23,841 (84.57)	66,437 (87.74)	174,834 (88.37)	264,583 (87.99)

* Liver includes intrahepatic bile duct.

https://doi.org/10.1371/journal.pone.0194407.t001 categories in our analyses. In this paper, both CCS diagnosis category descriptions and labels are used to represent various diagnoses. Since procedure codes were only available in a very small portion of records, we kept using ICD-9 or ICD-10 procedure codes without mapping. This study focuses on discovering patterns from seven types of cancer with high incident rates in New York State: rectum and anus cancer (15), liver and intrahepatic bile duct cancer (16), pancreas cancer (17), lung and bronchus cancer (19), breast cancer (24), prostate cancer (29) and Non-Hodgkin's lymphoma (38) [21]. There are 8,645,995 discharge records from 742,487 patients used in our work.

Data preparation

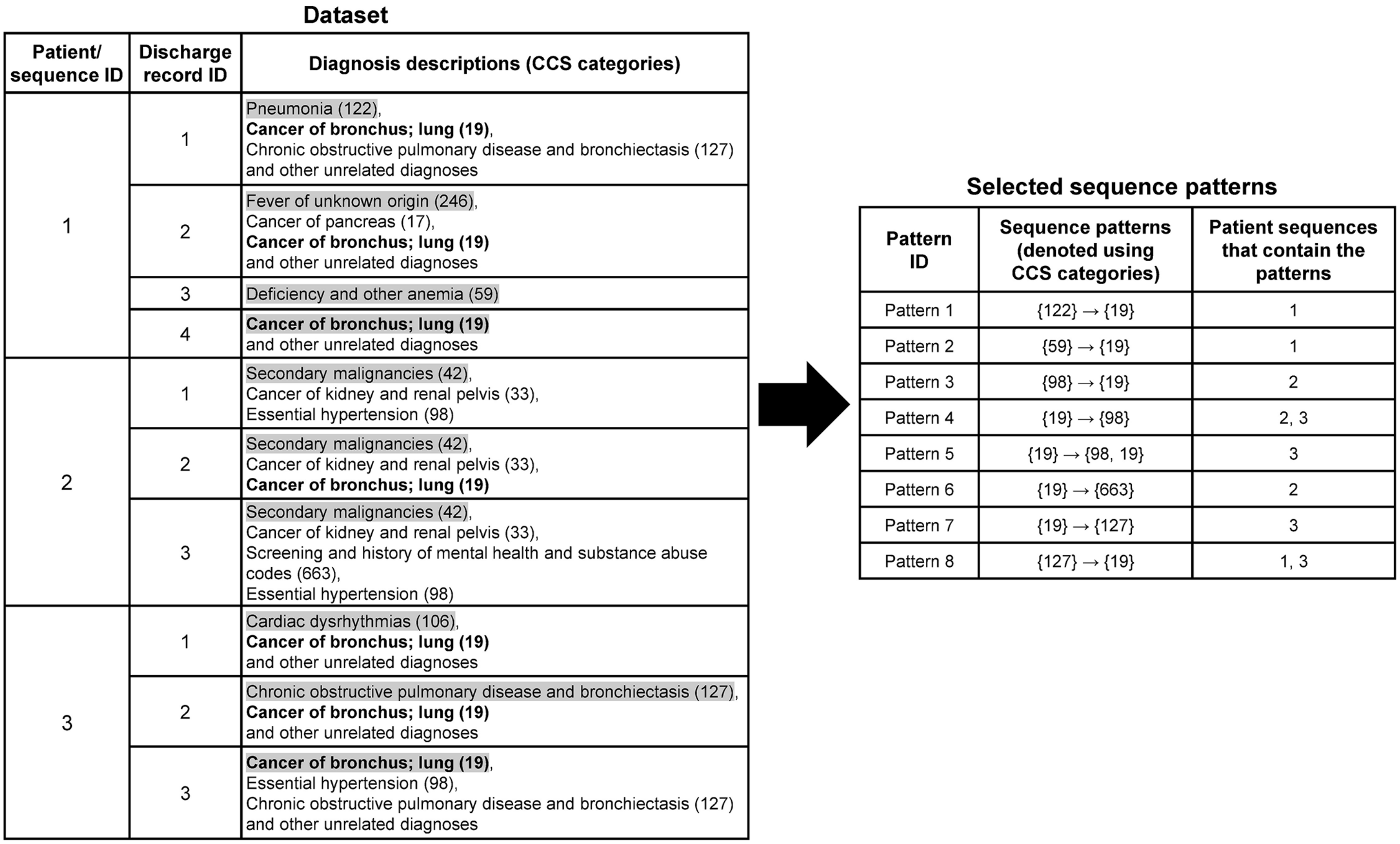
Each patient's discharge records were grouped together using an encrypted unique patient identifier in SPARCS and ordered by the corresponding admission dates. Discharge records containing AIDS/HIV or abortion diagnoses were removed from our analyses, because their admission dates and patient identifiers were redacted to comply with the Health Insurance Portability and Accountability Act (HIPAA) [22]. Patient level demographic information (i.e. age, gender, race and ethnicity) were collected from the first record of each patient. Patients were classified into cohorts having different types of cancer. A patient who had any cancer diagnosis was selected into the corresponding cohort. One patient could be in multiple cohorts because a person could have been diagnosed with different types of cancer. Table 1 shows the patient characteristics in our analyses. For each type of cancer, we studied the disparities of top 20 frequent co-occurrence and sequence patterns among different age groups (<=34, 35-54, 55-74 and >=75 years old) [23] and gender groups (male and female). We also analyzed disparities of co-occurrence patterns using discharge records from different claim types. Disparities among different race and ethnicity groups are not discussed in this paper, but we provide relevant results generated using discharge records from patients who had Non-Hodgkin's lymphoma (38) in S7 Table.

Patients' diagnosis sequences

For each patient, since discharge records were strictly ordered by admission dates, diagnosis information was also strictly ordered by corresponding admission dates. Thus, each patient's admission dates and diagnosis information constituted a diagnosis sequence. Each patient's diagnosis sequence was assigned a unique sequence ID, which was also an ID for this patient. Fig 1 shows a randomly selected example consisting of three diagnosis sequences from patients having lung and bronchus cancer (19), which is the targeted cancer in this example. Diagnoses are listed after the discharge record ID, a corresponding CCS category is marked in the parentheses following the description. The primary diagnosis in each discharge record is emphasized using grey shading. CCS category that represents the targeted cancer (i.e., lung and bronchus cancer) is highlighted in bold. In this example, pattern 5 ("{19}→{98, 19}") means that diagnosis "{cancer of bronchus; lung (19)}" happens before diagnoses "{essential hypertension (98), cancer of bronchus; lung (19)}". The former one diagnosis and the latter two diagnoses occur in different discharge records on different admission dates. And the latter two diagnoses occur on the same day.

Analysis methods

Apriori algorithm: Identifying frequent disease co-occurrence patterns. We adopted the Apriori algorithm [24] to discover frequent disease co-occurrence patterns and frequent procedure codes. The Apriori algorithm works by making multiple passes over the entire dataset and generating frequent co-occurrence itemsets (i.e., CCS categories on the same



Dataset

Patient/ sequence ID	Discharge record ID	Diagnosis descriptions (CCS categories)
1	1	Pneumonia (122), **Cancer of bronchus; lung (19)**, Chronic obstructive pulmonary disease and bronchiectasis (127) and other unrelated diagnoses
	2	Fever of unknown origin (246), Cancer of pancreas (17), **Cancer of bronchus; lung (19)** and other unrelated diagnoses
	3	Deficiency and other anemia (59)
	4	**Cancer of bronchus; lung (19)** and other unrelated diagnoses
2	1	Secondary malignancies (42), Cancer of kidney and renal pelvis (33), Essential hypertension (98)
	2	Secondary malignancies (42), Cancer of kidney and renal pelvis (33), **Cancer of bronchus; lung (19)**
	3	Secondary malignancies (42), Cancer of kidney and renal pelvis (33), Screening and history of mental health and substance abuse codes (663), Essential hypertension (98)
3	1	Cardiac dysrhythmias (106), **Cancer of bronchus; lung (19)** and other unrelated diagnoses
	2	Chronic obstructive pulmonary disease and bronchiectasis (127), **Cancer of bronchus; lung (19)** and other unrelated diagnoses
	3	**Cancer of bronchus; lung (19)**, Essential hypertension (98), Chronic obstructive pulmonary disease and bronchiectasis (127) and other unrelated diagnoses

Selected sequence patterns

Pattern ID	Sequence patterns (denoted using CCS categories)	Patient sequences that contain the patterns
Pattern 1	{122} → {19}	1
Pattern 2	{59} → {19}	1
Pattern 3	{98} → {19}	2
Pattern 4	{19} → {98}	2, 3
Pattern 5	{19} → {98, 19}	3
Pattern 6	{19} → {663}	2
Pattern 7	{19} → {127}	3
Pattern 8	{127} → {19}	1, 3

Fig 1. A random example of diagnosis sequences from patients having lung and bronchus cancer (19).

https://doi.org/10.1371/journal.pone.0194407.g001

admission date) by comparing their supports with a user-specified minimum support threshold. If the minimum support threshold is satisfied, this co-occurrence itemset is kept in the searching results; otherwise it is deleted from the searching results. In the first pass, the algorithm simply counts occurrences (i.e., support) of each CCS category and procedure code and determines which of them are large (i.e., satisfy the minimum support threshold). In each subsequent pass, there are two phases. First, the algorithm starts with large itemsets found in the previous pass to generate new potentially large itemsets, say candidate itemsets, by joining new CCS categories or procedure codes. Next, the dataset is scanned to calculate the support of each candidate itemset, and determine large itemsets in the current pass by comparing these supports with the minimum support threshold. Eventually, all co-occurrence itemsets containing disease codes or procedure codes and satisfy the minimum support threshold are generated.

Both primary and secondary CCS categories were used in the analysis of disease co-occurrence patterns. Only records containing targeted cancer CCS categories were selected. For instance, in the sequences illustrated in Fig 1, discharge records containing "Cancer of bronchus; lung (19)" were used, and diagnoses in the same discharge record would be included in a large itemset if they satisfied the minimum support. We discovered co-occurrence relationships not only between different diagnoses, but also between diagnoses and procedures. As for associations between cancers and procedures, since not all records contained valid procedure codes, we only chose records containing both targeted cancer CCS categories and valid procedure codes in this analyses. For each type of cancer, we selected the top 20 potentially meaningful co-occurrence itemsets that contained targeted cancer diagnoses as frequent co-occurrence patterns. We used apyori 1.1.1 [25], a Python package for Apriori algorithm to discover frequent co-occurrence patterns.

cSPADE algorithm: Discovering frequent disease sequence patterns. We used cSPADE, a frequent sequence mining algorithm [26] to discover frequent disease sequence patterns for different types of cancer. cSPADE algorithm generates frequent sequences iteratively based on a subsequence relation that if a sequence is frequent, then all subsequences of this sequence are also frequent [27]. In each iteration, the cSPADE algorithm also works by comparing the supports of candidate sequences with the minimum support threshold. It starts from the single item sequence, to sequences with maximal length by joining subsequences obtained from previous iteration. When computing the support of a subsequence, multiple occurrences of this subsequence in the same sequence are counted only once.

Fig 1 shows an example of diagnosis sequences containing both primary and secondary CCS categories where primary CCS categories are in grey shading. The length of a sequence pattern is the total number of itemsets in this sequence. For example, pattern 5, which means "{19}" happens before "{98, 19}", is a length-2 sequence pattern because there are two itemsets "{19}" and "{98, 19}" in this sequence. We set the minimum interval between two itemsets as one and the maximum interval as 180, such that the duration between admission dates of two consecutive itemsets in a sequence pattern must be within 1-180 days. That is, this algorithm can discover the association of two diagnoses happen within 180 days of each other. Previous studies found out that revisit intervals usually range from one month to over one year, and typical intervals are two, three or six months [28]. Thus, 180 days is an interval long enough to cover significant revisit diagnoses. For each type of cancer, we also kept the first 20 potentially meaningful subsequences that contained targeted cancer diagnoses as frequent sequence patterns. All frequent sequence patterns were mined using arulesSequences [29], a R package for cSPADE algorithm.

Statistical analyses. We selected top 20 co-occurrence patterns and top 20 sequence patterns to run statistical analyses for each type of cancer. Percentages of co-occurrence and sequence patterns were calculated and compared between age, gender, race, ethnicity group and claim type to evaluate disparities among these patient groups. For each of top co-occurrences as the dichotomous outcome, a generalized linear mixed-effect model was fit for the repeated measure data. Age, gender, race, ethnicity and claim type were all included in the same model as covariates. The within-subject dependence over repeated visits was adjusted using an unstructured covariance matrix. *p*-values based on *F*-tests were assessed to evaluate the overall significance of those covariates. For sequence patterns, multiple logistic regression models were fit for each sequence pattern as the dichotomous outcome. Similar to the analyses of co-occurrences, age, gender, race and ethnicity were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance. The Bonferroni's method was used to adjust *p*-values for multiple tests. All statistical analyses were performed using SAS v9.4 (the SAS Institute, Cary, NC) [30].

Results

We performed analyses on patients' diagnosis histories and focused on seven types of cancer with high incident rates in New York State. Meaningless results, such as patterns consisting of identical CCS categories, CCS categories that represent unspecific disease groups or serve administrative purposes, length-1 patterns and patterns irrelevant to targeted cancers, were ruled out from our analyses.

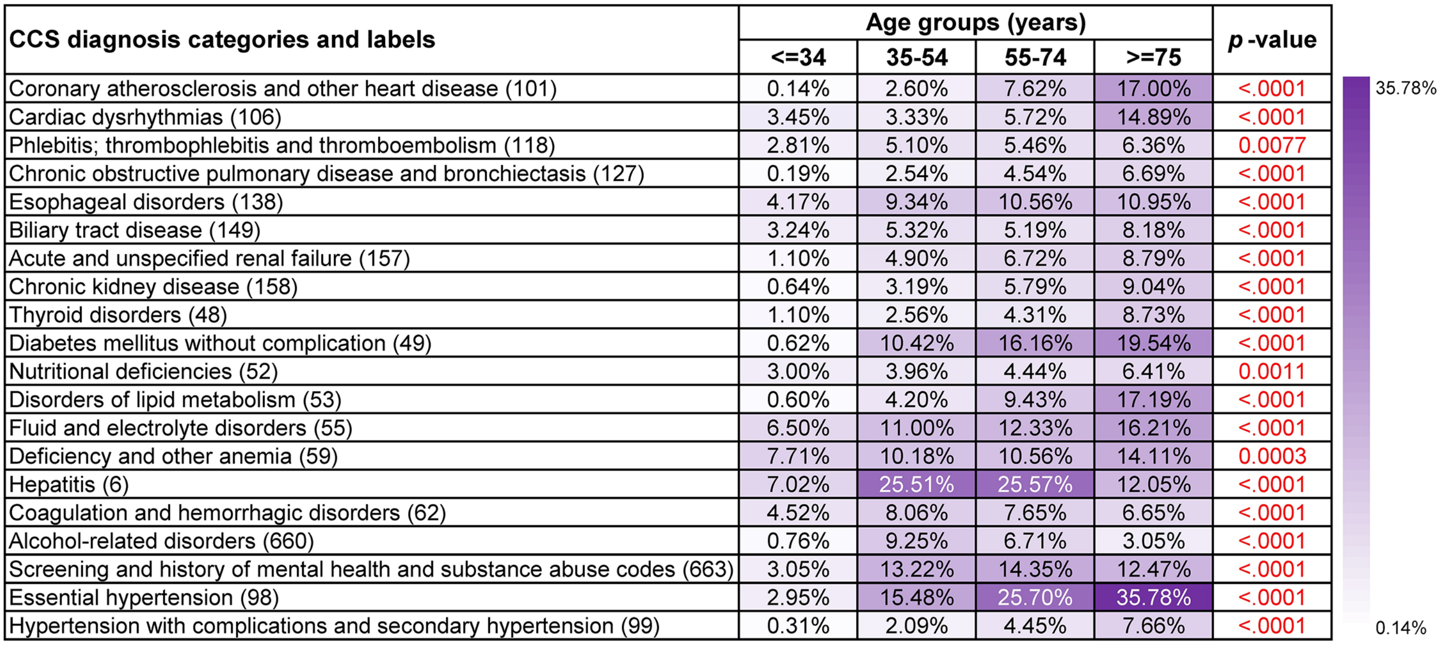
In this section, we mainly discuss patterns containing diagnoses closely related to the targeted cancers and present results of a few cancer types where significant pattern disparities are found. Other results are available in the supporting information.

Diagnosis co-occurrence patterns

We analyzed disparities of co-occurrence patterns among different age groups, gender groups and records from different claim types. Besides, we also discovered correlations between cancer diagnoses and procedures to see what procedures were frequently adopted to treat different types of cancer.

Frequent disease co-occurrence patterns in different age groups. Figs 2 and 3 show supports and *p*-values of top 20 frequent diagnoses that co-occurred with liver and intrahepatic bile duct cancer (16) and Non-Hodgkin's lymphoma (38), respectively. The *p*-values revealed that almost all of these diagnoses were significant with respect to age. Besides, supports of many top frequent diagnoses were the highest in the eldest age group (>=75 years old) and were the lowest in the youngest age group (<=34 years old).

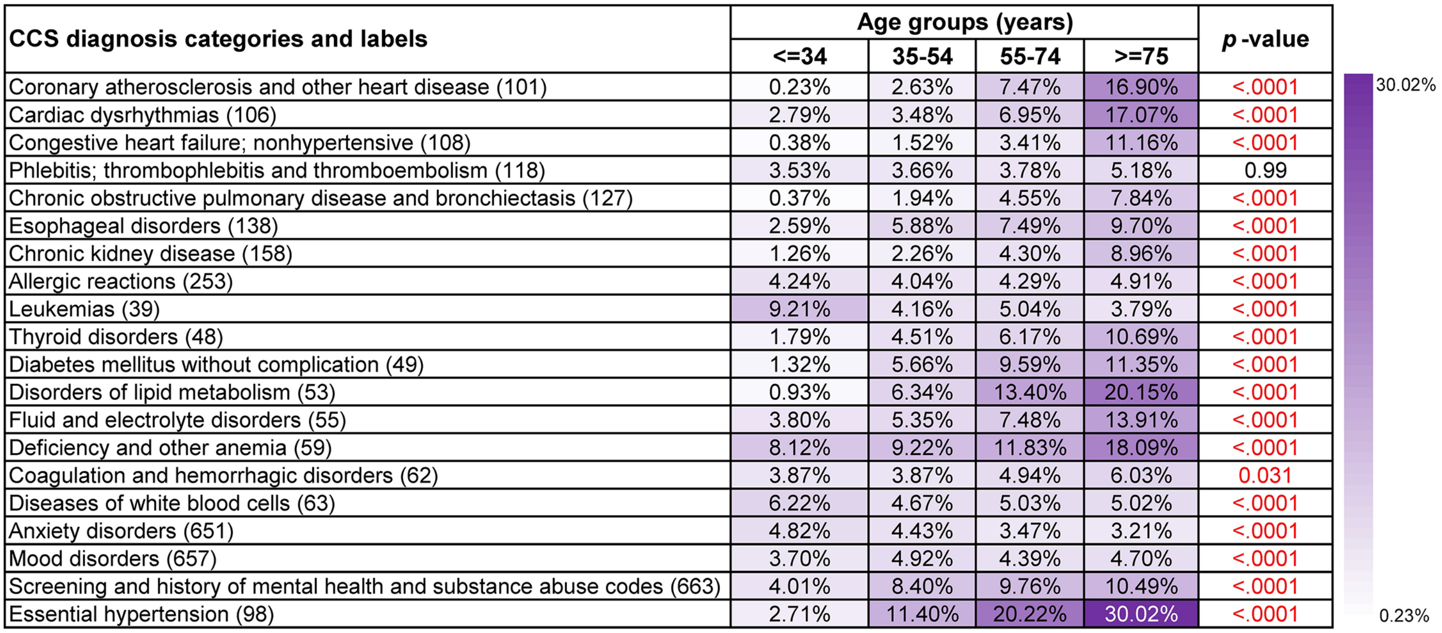
Fig 2 presents results of frequent co-occurrence patterns regarding liver and intrahepatic bile duct cancer (16). Among patients who were or under 34 years old, deficiency and other anemia (59) and hepatitis (6) were two most popular diagnosis co-occurrences (7.71% and 7.02%, respectively). Patients between 35-74 years old were also more likely to have hepatitis (6), while it was less seen among patients who were or over 75 years old (35-54 years old: 25.51%, 55-74 years old: 25.57%, >=75 years old: 12.05%, *p*-value<0.0001). Essential



CCS diagnosis categories and labels	Age groups (years)				*p* -value
	<=34	35-54	55-74	>=75	
Coronary atherosclerosis and other heart disease (101)	0.14%	2.60%	7.62%	17.00%	<.0001
Cardiac dysrhythmias (106)	3.45%	3.33%	5.72%	14.89%	<.0001
Phlebitis; thrombophlebitis and thromboembolism (118)	2.81%	5.10%	5.46%	6.36%	0.0077
Chronic obstructive pulmonary disease and bronchiectasis (127)	0.19%	2.54%	4.54%	6.69%	<.0001
Esophageal disorders (138)	4.17%	9.34%	10.56%	10.95%	<.0001
Biliary tract disease (149)	3.24%	5.32%	5.19%	8.18%	<.0001
Acute and unspecified renal failure (157)	1.10%	4.90%	6.72%	8.79%	<.0001
Chronic kidney disease (158)	0.64%	3.19%	5.79%	9.04%	<.0001
Thyroid disorders (48)	1.10%	2.56%	4.31%	8.73%	<.0001
Diabetes mellitus without complication (49)	0.62%	10.42%	16.16%	19.54%	<.0001
Nutritional deficiencies (52)	3.00%	3.96%	4.44%	6.41%	0.0011
Disorders of lipid metabolism (53)	0.60%	4.20%	9.43%	17.19%	<.0001
Fluid and electrolyte disorders (55)	6.50%	11.00%	12.33%	16.21%	<.0001
Deficiency and other anemia (59)	7.71%	10.18%	10.56%	14.11%	0.0003
Hepatitis (6)	7.02%	25.51%	25.57%	12.05%	<.0001
Coagulation and hemorrhagic disorders (62)	4.52%	8.06%	7.65%	6.65%	<.0001
Alcohol-related disorders (660)	0.76%	9.25%	6.71%	3.05%	<.0001
Screening and history of mental health and substance abuse codes (663)	3.05%	13.22%	14.35%	12.47%	<.0001
Essential hypertension (98)	2.95%	15.48%	25.70%	35.78%	<.0001
Hypertension with complications and secondary hypertension (99)	0.31%	2.09%	4.45%	7.66%	<.0001

Fig 2. Supports and *p*-values of the most frequent diagnoses that co-occurred with liver and intrahepatic bile duct cancer (16) in different age groups. Age, gender, race, ethnicity and claim type were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g002



CCS diagnosis categories and labels	Age groups (years)				*p*-value
	<=34	35-54	55-74	>=75	
Coronary atherosclerosis and other heart disease (101)	0.23%	2.63%	7.47%	16.90%	<.0001
Cardiac dysrhythmias (106)	2.79%	3.48%	6.95%	17.07%	<.0001
Congestive heart failure; nonhypertensive (108)	0.38%	1.52%	3.41%	11.16%	<.0001
Phlebitis; thrombophlebitis and thromboembolism (118)	3.53%	3.66%	3.78%	5.18%	0.99
Chronic obstructive pulmonary disease and bronchiectasis (127)	0.37%	1.94%	4.55%	7.84%	<.0001
Esophageal disorders (138)	2.59%	5.88%	7.49%	9.70%	<.0001
Chronic kidney disease (158)	1.26%	2.26%	4.30%	8.96%	<.0001
Allergic reactions (253)	4.24%	4.04%	4.29%	4.91%	<.0001
Leukemias (39)	9.21%	4.16%	5.04%	3.79%	<.0001
Thyroid disorders (48)	1.79%	4.51%	6.17%	10.69%	<.0001
Diabetes mellitus without complication (49)	1.32%	5.66%	9.59%	11.35%	<.0001
Disorders of lipid metabolism (53)	0.93%	6.34%	13.40%	20.15%	<.0001
Fluid and electrolyte disorders (55)	3.80%	5.35%	7.48%	13.91%	<.0001
Deficiency and other anemia (59)	8.12%	9.22%	11.83%	18.09%	<.0001
Coagulation and hemorrhagic disorders (62)	3.87%	3.87%	4.94%	6.03%	0.031
Diseases of white blood cells (63)	6.22%	4.67%	5.03%	5.02%	<.0001
Anxiety disorders (651)	4.82%	4.43%	3.47%	3.21%	<.0001
Mood disorders (657)	3.70%	4.92%	4.39%	4.70%	<.0001
Screening and history of mental health and substance abuse codes (663)	4.01%	8.40%	9.76%	10.49%	<.0001
Essential hypertension (98)	2.71%	11.40%	20.22%	30.02%	<.0001

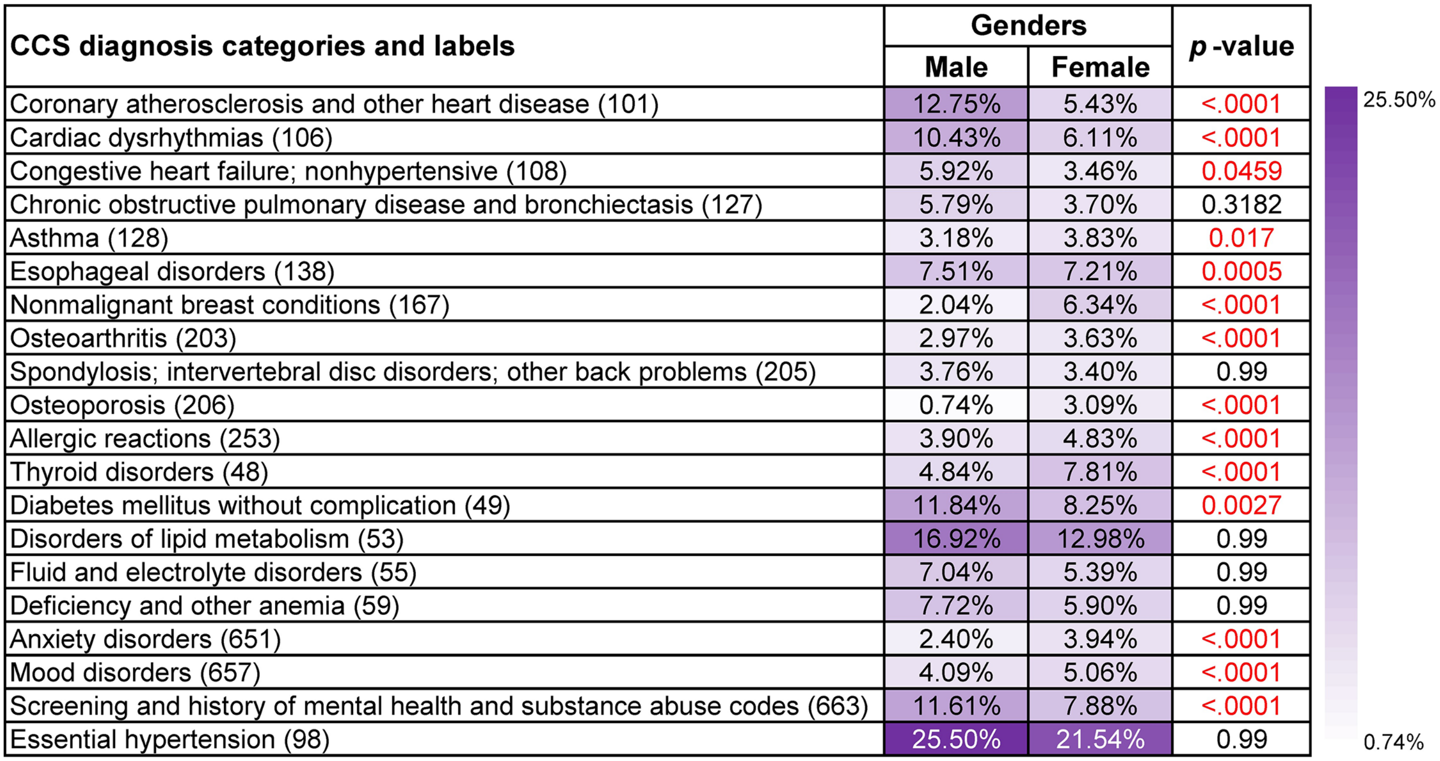
Fig 3. Supports and *p*-values of the most frequent diagnoses that co-occurred with Non-Hodgkin's lymphoma (38) in different age groups. Age, gender, race, ethnicity and claim type were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g003

hypertension (98) was another diagnosis co-occurrence that usually occurred among patients who were or over 55 years old (55-74 years old: 25.70%, >=75 years old: 35.78%, *p*-value < 0.0001). Results for Non-Hodgkin's lymphoma (38) were usually more representative and less noisy (Fig 3). For instance, leukemias (39) and diseases of white blood cells (63) were more frequent among patients who were or under 34 years old (9.21% and 6.22%, respectively). Essential hypertension (98) was also the most popular diagnosis co-occurrence with patients who were or over 55 years old (55-74 years old: 20.22%, >=75 years old: 30.02%).

Patients who had rectum and anus cancer (15) (S1 Table) were more frequently diagnosed with cancer of colon (14). However, this diagnosis was not significant with respect to age (*p*-value = 0.99). Biliary tract disease (149) and diabetes mellitus without complication (49) were significant diagnoses (*p*-value<0.0001) that co-occurred with pancreas cancer (17) (S3 Table). As for lung and bronchus cancer (19) (S4 Table), the frequent diagnosis most relevant to this cancer was chronic obstructive pulmonary disease and bronchiectasis (127), which ranked high on the list of the most frequent co-occurrence patterns for patients who were or over 55 years old (55-74 years old: 19.30%, >=75 years old: 26.01%, *p*-value<0.0001). Pneumonia (122) was another frequent diagnosis that was significant with respect to age (*p*-value<0.0001). Nonmalignant breast conditions (167) co-occurred more frequently with breast cancer (24) (S5 Table) among people who were or under 54 years old (<=34 years old: 7.06%, 35-54 years old: 8.40%). For co-occurrence patterns among patients who had prostate cancer (29) (S6 Table), genitourinary symptoms and ill-defined conditions (163) was a significant sign of patients who were or under 34 years old (12.27%).

Frequent disease co-occurrence patterns in different gender groups. The most frequent diagnosis co-occurrences for cancer types that are less gender specific, such as rectum and anus cancer (15) (S1 Table), liver and intrahepatic bile duct cancer (16) (S2 Table), pancreas cancer (17) (S3 Table) and lung and bronchus cancer (19) (S4 Table), demonstrated similar



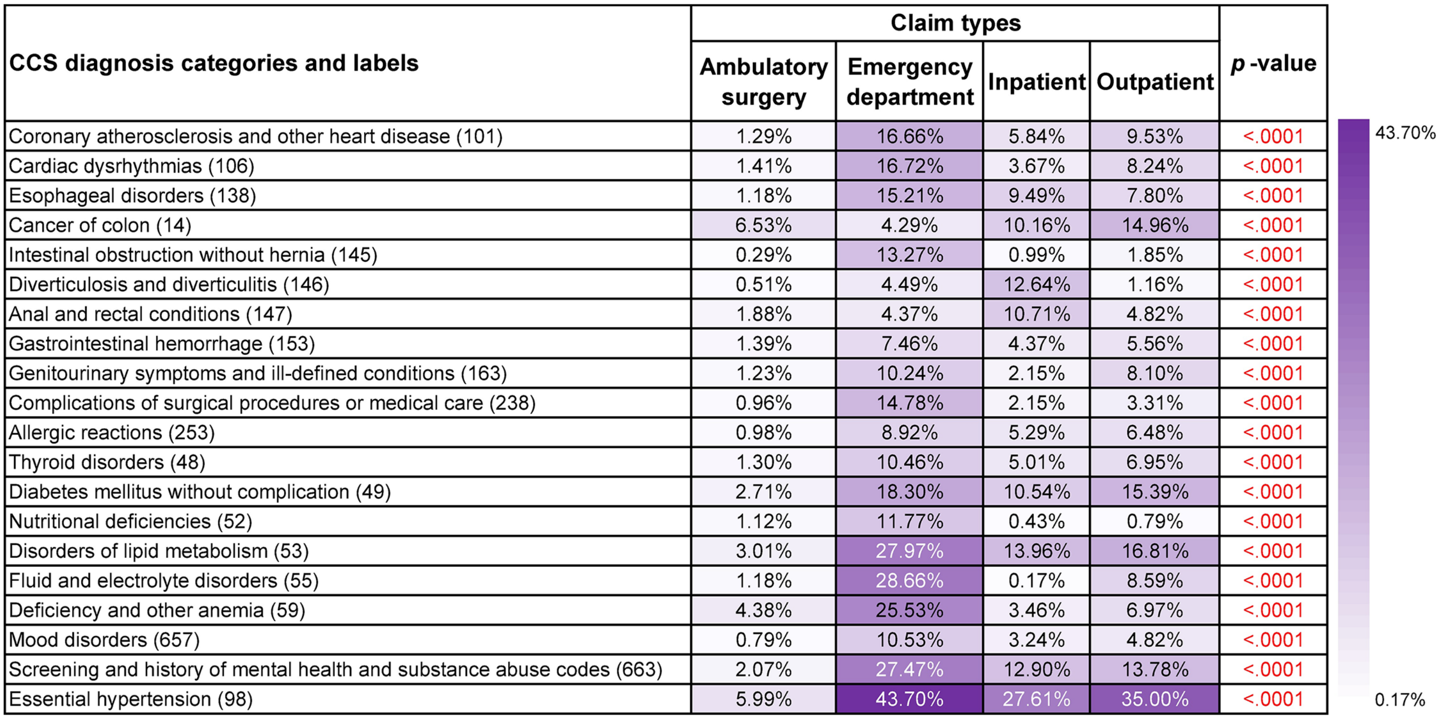
CCS diagnosis categories and labels	Genders		*p* -value
	Male	Female	
Coronary atherosclerosis and other heart disease (101)	12.75%	5.43%	<.0001
Cardiac dysrhythmias (106)	10.43%	6.11%	<.0001
Congestive heart failure; nonhypertensive (108)	5.92%	3.46%	0.0459
Chronic obstructive pulmonary disease and bronchiectasis (127)	5.79%	3.70%	0.3182
Asthma (128)	3.18%	3.83%	0.017
Esophageal disorders (138)	7.51%	7.21%	0.0005
Nonmalignant breast conditions (167)	2.04%	6.34%	<.0001
Osteoarthritis (203)	2.97%	3.63%	<.0001
Spondylosis; intervertebral disc disorders; other back problems (205)	3.76%	3.40%	0.99
Osteoporosis (206)	0.74%	3.09%	<.0001
Allergic reactions (253)	3.90%	4.83%	<.0001
Thyroid disorders (48)	4.84%	7.81%	<.0001
Diabetes mellitus without complication (49)	11.84%	8.25%	0.0027
Disorders of lipid metabolism (53)	16.92%	12.98%	0.99
Fluid and electrolyte disorders (55)	7.04%	5.39%	0.99
Deficiency and other anemia (59)	7.72%	5.90%	0.99
Anxiety disorders (651)	2.40%	3.94%	<.0001
Mood disorders (657)	4.09%	5.06%	<.0001
Screening and history of mental health and substance abuse codes (663)	11.61%	7.88%	<.0001
Essential hypertension (98)	25.50%	21.54%	0.99

Fig 4. Supports and *p*-values of the most frequent diagnoses that co-occurred with breast cancer (24) in different gender groups. Age, gender, race, ethnicity and claim type were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g004

trends in males and females with only very few disparities. For example, liver and intrahepatic cancer (16), females were usually at a higher risk of having deficiency and other anemia (59) than males (male: 10.32%, female: 12.14%, *p*-value<0.0001). Males were more likely to have hepatitis (6) compared with females (male: 25.69%, female: 16.83%, *p*-value<0.0001). Thyroid disorders (48) was more popular with females among patients who had Non-Hodgkin's lymphoma (38) (S7 Table) and breast cancer (24) (Fig 4), but was not high on the list of the top frequent disease co-occurrences for males (*p*-values<0.0001). It can also be observed that heart disease like coronary atherosclerosis (101) affected males more than females across all seven types of cancer (*p*-values<0.0001).

Frequent co-occurrence patterns from different claim types. Distribution of frequent diagnosis co-occurrence patterns regarding claim types differed among all seven types of cancer (*p*-values<0.0001). For instance, colon cancer (14) were the most frequent diagnosis only in ambulatory surgery visits from patients having rectum and anus cancer (15) (Fig 5), biliary tract disease (149) were comparatively frequent in ambulatory surgery visits and inpatient hospital stays with respect to pancreas cancer (17) (S3 Table). However, some common patterns can still be identified. Most of the discharge records of lung and bronchus cancer (19) (S4 Table), prostate cancer (29) (S6 Table) and Non-Hodgkin's lymphoma (38) (S7 Table) came from ambulatory surgery visits, least of them were from inpatient care. Most discharge records for rectum and anus cancer (15) (Fig 5) and liver and intrahepatic bile duct cancer (16) (S2 Table) were collected from emergency department visits.

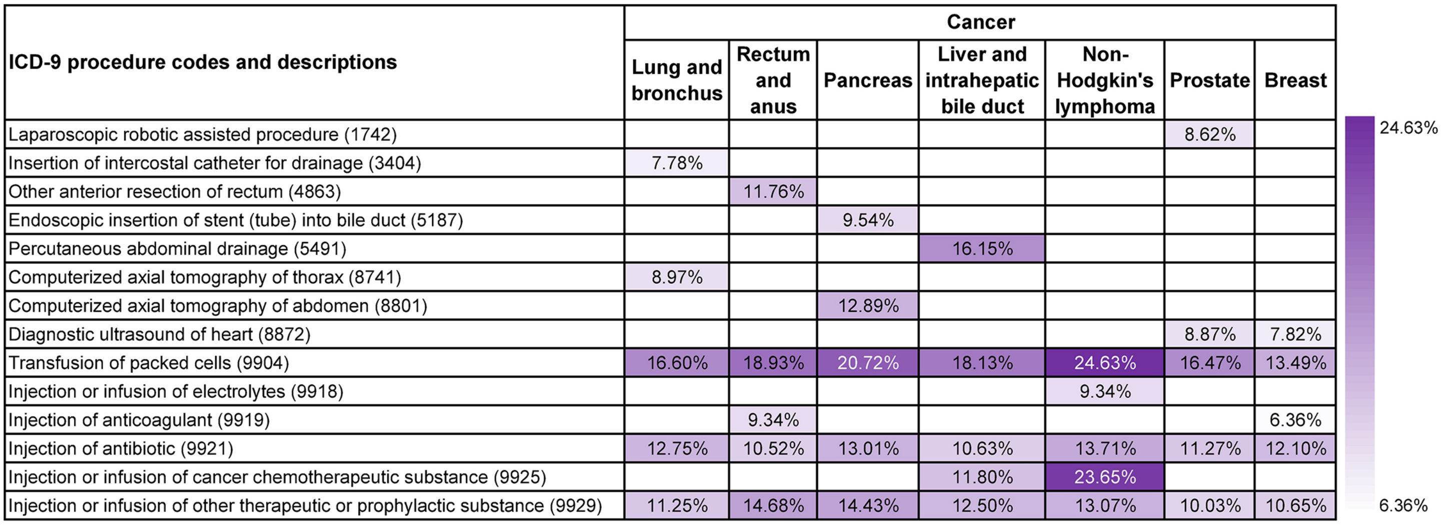


CCS diagnosis categories and labels	Claim types				*p* -value
	Ambulatory surgery	Emergency department	Inpatient	Outpatient	
Coronary atherosclerosis and other heart disease (101)	1.29%	16.66%	5.84%	9.53%	<.0001
Cardiac dysrhythmias (106)	1.41%	16.72%	3.67%	8.24%	<.0001
Esophageal disorders (138)	1.18%	15.21%	9.49%	7.80%	<.0001
Cancer of colon (14)	6.53%	4.29%	10.16%	14.96%	<.0001
Intestinal obstruction without hernia (145)	0.29%	13.27%	0.99%	1.85%	<.0001
Diverticulosis and diverticulitis (146)	0.51%	4.49%	12.64%	1.16%	<.0001
Anal and rectal conditions (147)	1.88%	4.37%	10.71%	4.82%	<.0001
Gastrointestinal hemorrhage (153)	1.39%	7.46%	4.37%	5.56%	<.0001
Genitourinary symptoms and ill-defined conditions (163)	1.23%	10.24%	2.15%	8.10%	<.0001
Complications of surgical procedures or medical care (238)	0.96%	14.78%	2.15%	3.31%	<.0001
Allergic reactions (253)	0.98%	8.92%	5.29%	6.48%	<.0001
Thyroid disorders (48)	1.30%	10.46%	5.01%	6.95%	<.0001
Diabetes mellitus without complication (49)	2.71%	18.30%	10.54%	15.39%	<.0001
Nutritional deficiencies (52)	1.12%	11.77%	0.43%	0.79%	<.0001
Disorders of lipid metabolism (53)	3.01%	27.97%	13.96%	16.81%	<.0001
Fluid and electrolyte disorders (55)	1.18%	28.66%	0.17%	8.59%	<.0001
Deficiency and other anemia (59)	4.38%	25.53%	3.46%	6.97%	<.0001
Mood disorders (657)	0.79%	10.53%	3.24%	4.82%	<.0001
Screening and history of mental health and substance abuse codes (663)	2.07%	27.47%	12.90%	13.78%	<.0001
Essential hypertension (98)	5.99%	43.70%	27.61%	35.00%	<.0001

Fig 5. Supports and *p*-values of the most frequent diagnoses that co-occurred with rectum and anus cancer (15) in discharge records for different claim types. Age, gender, race, ethnicity and claim type were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g005

Frequent co-occurrences of cancer diagnoses and procedure codes. Fig 6 illustrates the procedure codes that co-occurred most frequently with different cancer diagnoses. We used discharge records that contained both targeted cancer diagnoses and valid procedure codes in this study. We also selected the top 20 most frequent procedure codes for each type of cancer, while only present top five most frequent procedure codes here. Transfusion of packed cells (9904), injection of antibiotic (9921) and injection or infusion of other therapeutic or prophylactic substance (9929) appeared in the top five most frequent procedure codes of all seven types of cancer and were usually in the first three places. Moreover, transfusion of packed cells (9904) always ranked the first for each type of cancer. Pattern disparities among different types of cancer could also be identified. Insertion of intercostal catheter for drainage (3404) (7.78%) and computerized axial tomography of thorax (8741) (8.97%) were popular regarding lung and bronchus cancer (19). Other anterior resection of rectum (4863) (11.76%) was more common in treating rectum and anus cancer (15). Computerized axial tomography of abdomen (8801) (12.89%) and endoscopic insertion of stent (tube) into bile duct (5187) (9.54%) were more frequent in results for pancreas cancer (17). Percutaneous abdominal drainage (5491) (16.15%) was popular in treating liver and intrahepatic bile duct cancer (16). Injection or infusion of electrolytes (9918) (9.34%) was more frequently used to treat Non-Hodgkin's lymphoma (38). Laparoscopic robotic assisted procedure (1742) (8.62%) was only found in top five frequent patterns for prostate cancer (29). Injection or infusion of cancer chemotherapeutic substance (9925) was most frequently used to treat liver and intrahepatic bile duct cancer (16) (11.80%) and Non-Hodgkin's lymphoma (38) (23.65%).



ICD-9 procedure codes and descriptions	Cancer						
	Lung and bronchus	Rectum and anus	Pancreas	Liver and intrahepatic bile duct	Non-Hodgkin's lymphoma	Prostate	Breast
Laparoscopic robotic assisted procedure (1742)						8.62%	
Insertion of intercostal catheter for drainage (3404)	7.78%						
Other anterior resection of rectum (4863)		11.76%					
Endoscopic insertion of stent (tube) into bile duct (5187)			9.54%				
Percutaneous abdominal drainage (5491)				16.15%			
Computerized axial tomography of thorax (8741)	8.97%						
Computerized axial tomography of abdomen (8801)			12.89%				
Diagnostic ultrasound of heart (8872)						8.87%	7.82%
Transfusion of packed cells (9904)	16.60%	18.93%	20.72%	18.13%	24.63%	16.47%	13.49%
Injection or infusion of electrolytes (9918)					9.34%		
Injection of anticoagulant (9919)		9.34%					6.36%
Injection of antibiotic (9921)	12.75%	10.52%	13.01%	10.63%	13.71%	11.27%	12.10%
Injection or infusion of cancer chemotherapeutic substance (9925)				11.80%	23.65%		
Injection or infusion of other therapeutic or prophylactic substance (9929)	11.25%	14.68%	14.43%	12.50%	13.07%	10.03%	10.65%

Fig 6. Supports of the most frequent procedures used to treat different types of cancer.

https://doi.org/10.1371/journal.pone.0194407.g006

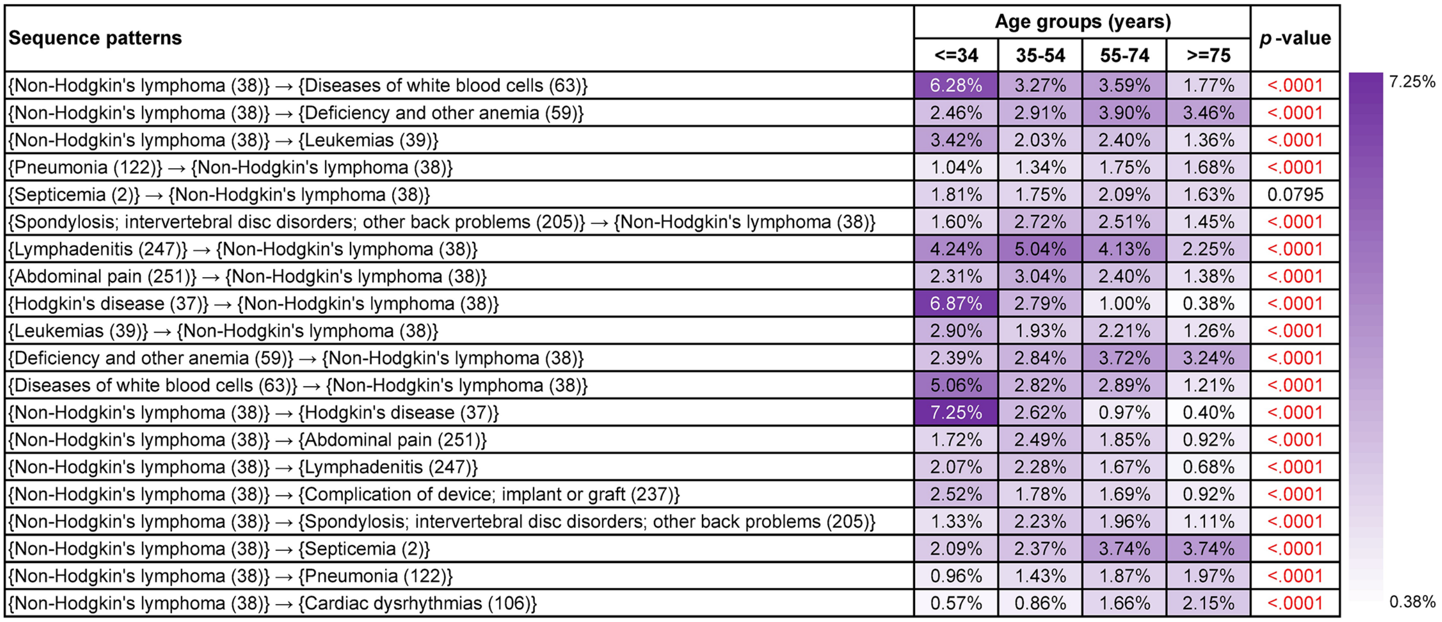
Diagnosis sequence patterns

In our analyses, we searched on full patient diagnosis sequences using primary CCS categories only, because primary diagnoses could help detect more clinically meaningful patterns [31]. We also used both primary and secondary diagnoses and ran analyses on sequences from patients who had Non-Hodgkin's lymphoma (38). Results are available in S7 Table. By comparing results presented in Figs 3 and 7 and S7 Table, we found that although a combination of primary and secondary CCS categories contained richer diagnosis information, the information could be redundant and noisy.

We only present length-2 diagnosis sequence patterns in this section, as longer sequences in our analyses usually consisted of repeated CCS categories representing follow-up visits rather than disease progression. Moreover, since only primary diagnoses were used in mining sequence patterns and patients having targeted diagnoses usually did not have these CCS categories as the primary diagnoses, supports of patterns presented in this section are comparatively lower than patterns generated using both primary and secondary diagnosis information (S7 Table).

Frequent sequence patterns in different age groups. Diagnoses in sequence patterns were usually more closely correlated with targeted cancers than those in co-occurrence patterns.

Non-Hodgkin's lymphoma (38) was a typical cancer type where significant disparities with respect to age were observed. Fig 7 presents results of frequent sequence patterns for Non-Hodgkin's lymphoma (38) in different age groups. Hodgkin's disease (37), leukemias (39), diseases of white blood cells (63) and lymphadenitis (247) were four essential comorbidities. "{Non-Hodgkin's lymphoma (38)}→{diseases of white blood cells (63)}" (<=34 years old: 6.28%, *p*-value < 0.0001), "{Hodgkin's disease (37)}→{Non-Hodgkin's lymphoma (38)}" (<=34 years old: 6.87%, *p*-value<0.0001), "{Non-Hodgkin's lymphoma (38)}→{Hodgkin's disease (37)}" (<=34 years old: 7.25%, *p*-value<0.0001), "{diseases of white blood cells (63)}→{Non-Hodgkin's lymphoma (38)}" (<=34 years old: 5.06%, *p*-value<0.0001) were four sequence patterns with the highest supports discovered from patients who were or under 34



Sequence patterns	Age groups (years)				*p* -value
	<=34	35-54	55-74	>=75	
{Non-Hodgkin's lymphoma (38)} → {Diseases of white blood cells (63)}	6.28%	3.27%	3.59%	1.77%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Deficiency and other anemia (59)}	2.46%	2.91%	3.90%	3.46%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Leukemias (39)}	3.42%	2.03%	2.40%	1.36%	<.0001
{Pneumonia (122)} → {Non-Hodgkin's lymphoma (38)}	1.04%	1.34%	1.75%	1.68%	<.0001
{Septicemia (2)} → {Non-Hodgkin's lymphoma (38)}	1.81%	1.75%	2.09%	1.63%	0.0795
{Spondylosis; intervertebral disc disorders; other back problems (205)} → {Non-Hodgkin's lymphoma (38)}	1.60%	2.72%	2.51%	1.45%	<.0001
{Lymphadenitis (247)} → {Non-Hodgkin's lymphoma (38)}	4.24%	5.04%	4.13%	2.25%	<.0001
{Abdominal pain (251)} → {Non-Hodgkin's lymphoma (38)}	2.31%	3.04%	2.40%	1.38%	<.0001
{Hodgkin's disease (37)} → {Non-Hodgkin's lymphoma (38)}	6.87%	2.79%	1.00%	0.38%	<.0001
{Leukemias (39)} → {Non-Hodgkin's lymphoma (38)}	2.90%	1.93%	2.21%	1.26%	<.0001
{Deficiency and other anemia (59)} → {Non-Hodgkin's lymphoma (38)}	2.39%	2.84%	3.72%	3.24%	<.0001
{Diseases of white blood cells (63)} → {Non-Hodgkin's lymphoma (38)}	5.06%	2.82%	2.89%	1.21%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Hodgkin's disease (37)}	7.25%	2.62%	0.97%	0.40%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Abdominal pain (251)}	1.72%	2.49%	1.85%	0.92%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Lymphadenitis (247)}	2.07%	2.28%	1.67%	0.68%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Complication of device; implant or graft (237)}	2.52%	1.78%	1.69%	0.92%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Spondylosis; intervertebral disc disorders; other back problems (205)}	1.33%	2.23%	1.96%	1.11%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Septicemia (2)}	2.09%	2.37%	3.74%	3.74%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Pneumonia (122)}	0.96%	1.43%	1.87%	1.97%	<.0001
{Non-Hodgkin's lymphoma (38)} → {Cardiac dysrhythmias (106)}	0.57%	0.86%	1.66%	2.15%	<.0001

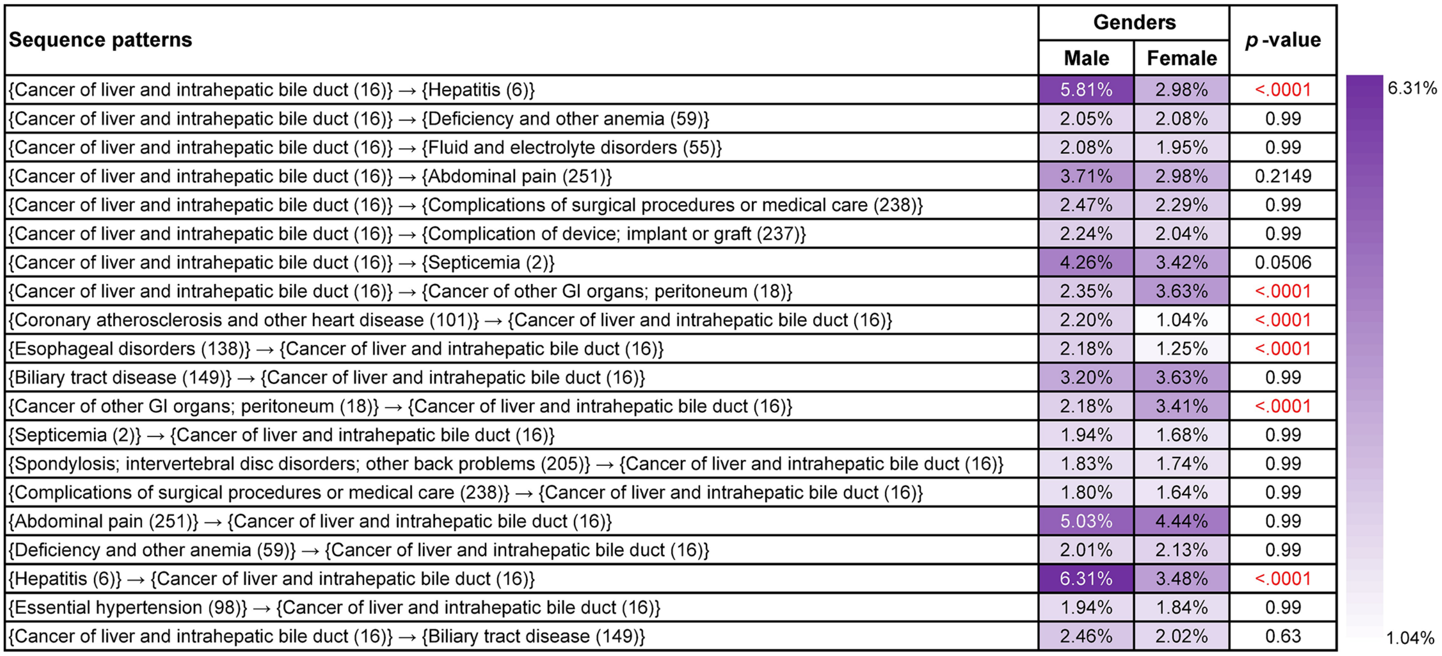
Fig 7. Supports and *p*-values of the most frequent sequence patterns found from patients who had Non-Hodgkin's lymphoma (38) in different age groups. Age, gender, race and ethnicity were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g007

years old. Besides, "{lymphadenitis (247)}→{Non-Hodgkin's lymphoma (38)}" (35-54 years old: 5.04%, 55-74 years old: 4.13%, *p*-value<0.0001) was the most frequent sequence pattern among patients between 35-74 years old.

Patients who had rectum and anus cancer (15) (S1 Table) were more frequently diagnosed with colon cancer (14). Sequence patterns containing colon cancer (14) were the most common ones across all age groups (*p*-values<0.0001). Patients between 0-54 years old had pattern "{cancer of colon (14)}→{cancer of rectum and anus (15)}" (<=34 years old: 7.68%, 35-54 years old: 11.94%) with supports slightly lower than its reversed sequence "{cancer of rectum and anus (15)}→{cancer of colon (14)}" (<=34 years old: 8.56%, 35-54 years old: 12.27%). However, the picture was different among patients who were or over 55 years old: the former (55-74 years old: 10.93%, >=75 years old: 6.66%) was more frequent than the latter (55-74 years old: 10.90%, >=75 years old: 6.28%). Also, patients within this age group (>=55 years old) were more likely to had pattern "{cancer of rectum and anus (15)}→{septicemia (2)}" (55-74 years old: 2.52%, >=75 years old: 2.36%). Sequences containnig anal and rectal conditions (147) were more common among patients who were or under 54 years old (*p*-values<0.0001).

Patterns regarding liver and intrahepatic bile duct cancer (16) (S2 Table) also varied among different age groups. Hepatitis (6) and abdominal pain (251) were two significant diagnoses in those patterns. Sequence "{cancer of liver and intrahepatic bile duct (16)}→{hepatitis (6)}" (35-54 yeas old: 5.75%, 55-74 years old: 5.73% years old, *p*-value<0.0001) and sequence "{hepatitis (6)}→{cancer of liver and intrahepatic bile duct (16)}" (35-54 yeas old: 6.95%, 55-74 years old: 6.14% years old, *p*-value<0.0001) were more common among patients who were 35-74 years old. Pattern "{cancer of liver and intrahepatic bile duct (16)}→{abdominal pain (251)}" (35-54 years old: 4.95%, *p*-value<0.0001) and "{abdominal pain (251)}→{cancer of liver and intrahepatic bile duct (16)}" (35-54 years old: 7.07%, *p*-value<0.0001) were more frequent among patients between 35-54 years old.



Sequence patterns	Genders		*p*-value
	Male	Female	
{Cancer of liver and intrahepatic bile duct (16)} → {Hepatitis (6)}	5.81%	2.98%	<.0001
{Cancer of liver and intrahepatic bile duct (16)} → {Deficiency and other anemia (59)}	2.05%	2.08%	0.99
{Cancer of liver and intrahepatic bile duct (16)} → {Fluid and electrolyte disorders (55)}	2.08%	1.95%	0.99
{Cancer of liver and intrahepatic bile duct (16)} → {Abdominal pain (251)}	3.71%	2.98%	0.2149
{Cancer of liver and intrahepatic bile duct (16)} → {Complications of surgical procedures or medical care (238)}	2.47%	2.29%	0.99
{Cancer of liver and intrahepatic bile duct (16)} → {Complication of device; implant or graft (237)}	2.24%	2.04%	0.99
{Cancer of liver and intrahepatic bile duct (16)} → {Septicemia (2)}	4.26%	3.42%	0.0506
{Cancer of liver and intrahepatic bile duct (16)} → {Cancer of other GI organs; peritoneum (18)}	2.35%	3.63%	<.0001
{Coronary atherosclerosis and other heart disease (101)} → {Cancer of liver and intrahepatic bile duct (16)}	2.20%	1.04%	<.0001
{Esophageal disorders (138)} → {Cancer of liver and intrahepatic bile duct (16)}	2.18%	1.25%	<.0001
{Biliary tract disease (149)} → {Cancer of liver and intrahepatic bile duct (16)}	3.20%	3.63%	0.99
{Cancer of other GI organs; peritoneum (18)} → {Cancer of liver and intrahepatic bile duct (16)}	2.18%	3.41%	<.0001
{Septicemia (2)} → {Cancer of liver and intrahepatic bile duct (16)}	1.94%	1.68%	0.99
{Spondylosis; intervertebral disc disorders; other back problems (205)} → {Cancer of liver and intrahepatic bile duct (16)}	1.83%	1.74%	0.99
{Complications of surgical procedures or medical care (238)} → {Cancer of liver and intrahepatic bile duct (16)}	1.80%	1.64%	0.99
{Abdominal pain (251)} → {Cancer of liver and intrahepatic bile duct (16)}	5.03%	4.44%	0.99
{Deficiency and other anemia (59)} → {Cancer of liver and intrahepatic bile duct (16)}	2.01%	2.13%	0.99
{Hepatitis (6)} → {Cancer of liver and intrahepatic bile duct (16)}	6.31%	3.48%	<.0001
{Essential hypertension (98)} → {Cancer of liver and intrahepatic bile duct (16)}	1.94%	1.84%	0.99
{Cancer of liver and intrahepatic bile duct (16)} → {Biliary tract disease (149)}	2.46%	2.02%	0.63

Fig 8. Supports and *p*-values of the most frequent sequence patterns found from patients who had liver and intrahepatic bile duct cancer (16) in different gender groups. Age, gender, race and ethnicity were treated as covariates and *p*-values based on *F*-tests were used to assess their overall significance.

https://doi.org/10.1371/journal.pone.0194407.g008

Frequent sequence patterns in different gender groups. Similar to disease co-occurrence patterns, distribution of top frequent sequence patterns for cancer types that are less gender specific, such as rectum and anus (15) (S1 Table), pancreas cancer (17) (S3 Table), lung and bronchus cancer (19) (S4 Table) and Non-Hodgkin's lymphoma (38) (S7 Table) demonstrated similar trends for both genders, but supports for males were usually higher than females. However, liver and intrahepatic bile duct cancer (16) was an exception. As shown in Fig 8, female patients were more likely to have cancer of other GI organs or peritoneum (18). These patterns were "{cancer of liver and intrahepatic bile duct (16)}→{cancer of other GI organs or peritoneum (18)}" (male: 2.35%, female: 3.63%, *p*-value<0.0001), "{cancer of other GI organs or peritoneum (18)}→{cancer of liver and intrahepatic bile duct (16)}" (male: 2.18%, female: 3.41%, *p*-value<0.0001). Although breast cancer (24) occured more commonly in females than in males, the trends of frequent sequence patterns were similar between different gender groups (S5 Table). For instance, "{nonmalignant breast conditions (167)}→{cancer of breast (24)}" (male: 3.86%, female: 8.00%, *p*-value<0.0001) and "{cancer of breast (24)}→{nonmalignant breast conditions (167)}" (male: 1.58%, female: 5.23%, *p*-value<0.0001) were highly popular with both genders.

Discussion

Although our work focused on discoveries of patterns and disparities in different patient groups, there were many common diagnoses appearing in almost results of all seven types of cancer, especially in diagnosis co-occurrence patterns. In co-occurrence patterns, patients who were and over 75 years old had much higher risk having cardiovascular diseases such as coronary atherosclerosis (101) and cardiac dysrhythmias (106). Essential hypertension (98) was another popular diagnosis with elder patients and it was high on the list of the top 20 frequent co-occurrence patterns of every cancer. Besides, disorders of lipid metabolism (53), fluid and electrolyte disorders (55), diabetes mellitus without complications (49) and deficiency and other anemia (59) were also common diagnoses that co-occurred frequently with different cancer diagnoses.

In our study, we used both primary and secondary diagnoses to discover disease co-occurrence patterns and only primary diagnoses to identify sequence patterns. As aforementioned, primary diagnosis codes are the main reason for a hospital visit and secondary diagnosis codes represent conditions that co-exist during the same hospital visit or stay. The sequence patterns usually contained diagnoses that were highly correlated with each targeted cancer, but many of these diagnoses were not available in the most frequent co-occurrence patterns. Thus, primary diagnosis information was more accurate and precise, and secondary diagnosis codes provided richer but redundant and noisy information. Moreover, sequence patterns conveyed information that were time dependent. For example, the support of a sequence was usually different from the support of its reversed sequence. This phenomenon might weigh significantly in studying disease progression.

Conclusions

Open data initiatives make large scale healthcare data available and provide us a unique opportunity for discovering patterns using data mining methods. We adopted Apriori algorithm and cSPADE algorithm to discover frequent disease co-occurrence and sequence patterns among cancer patients in New York State using SPARCS data. We studied seven types of cancer with high incident rates in New York State and focused on disparities of diagnosis co-occurrence patterns and diagnosis sequence patterns from patients' diagnosis histories with respect to age, gender as well as claim types. Our results suggest that the methods can generate potentially interesting and clinically meaningful disease co-occurrence and sequence patterns, which can be used to study comorbidities and disease progression for improving the management of multiple diseases of cancer patients.

Supporting information

S1 Table. Other results for rectum and anus cancer (15).
(XLSX)

S2 Table. Other results for liver and intrahepatic bile duct cancer (16).
(XLSX)

S3 Table. Other results for pancreas cancer (17).
(XLSX)

S4 Table. Other results for lung and bronchus cancer (19).
(XLSX)

S5 Table. Other results for breast cancer (24).
(XLSX)

S6 Table. Other results for prostate cancer (29).
(XLSX)

S7 Table. Other results for Non-Hodgkin's lymphoma (38).
(XLSX)

Author Contributions

Conceptualization: Yu Wang, Fusheng Wang.

Data curation: Yu Wang.

Formal analysis: Yu Wang, Wei Hou.

Funding acquisition: Fusheng Wang.

Investigation: Yu Wang.

Methodology: Yu Wang, Fusheng Wang.

Project administration: Fusheng Wang.

Resources: Fusheng Wang.

Software: Yu Wang.

Supervision: Fusheng Wang.

Validation: Yu Wang.

Visualization: Yu Wang.

Writing – original draft: Yu Wang, Wei Hou.

Writing – review & editing: Yu Wang, Wei Hou, Fusheng Wang.

References

1. Munson ME, Wrobel JS, Holmes CM, Hanauer DA. Data mining for identifying novel associations and temporal relationships with Charcot foot. Journal of diabetes research. 2014 Apr 27; 2014. https://doi.org/10.1155/2014/214353 PMID: 24868558
2. Kost R, Littenberg B, Chen ES. Exploring generalized association rule mining for disease co-occurrences. In AMIA Annual Symposium Proceedings 2012 (Vol. 2012, p. 1284). American Medical Informatics Association.
3. Wang F, Lee N, Hu J, Sun J, Ebadollahi S, Laine AF. A framework for mining signatures from event sequences and its applications in healthcare data. IEEE transactions on pattern analysis and machine intelligence. 2013 Feb; 35(2):272–85. https://doi.org/10.1109/TPAMI.2012.111 PMID: 22585098
4. Klema J, Nováková L, Karel F, Stepankova O, Zelezny F. Sequential data mining: A comparative case study in development of atherosclerosis risk factors. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2008 Jan; 38(1):3–15. https://doi.org/10.1109/TSMCC.2007.906055
5. López-Soto PJ, Smolensky MH, Sackett-Lundeen LL, De Giorgi A, Rodríguez-Borrego MA, Manfredini R, et al. Temporal Patterns of In-Hospital Falls of Elderly Patients. Nursing Research. 2016 Nov 1; 65 (6):435–45. https://doi.org/10.1097/NNR.0000000000000184 PMID: 27801714
6. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature communications. 2014 Jun 24; 5. https://doi.org/10.1038/ncomms5022
7. Tsoi AC, Zhang S, Hagenbuchner M. Pattern discovery on Australian medical claim data-a systematic approach. IEEE transactions on knowledge and data engineering. 2005 Oct; 17(10):1420–35. https://doi.org/10.1109/TKDE.2005.168
8. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE journal of biomedical and health informatics. 2015 Jul; 19(4):1216–23. https://doi.org/10.1109/JBHI.2015.2404829 PMID: 25706935
9. New York State Department of Health. Statewide Planning and Research Cooperative System (SPARCS). 2016. Available from: https://www.health.ny.gov/statistics/sparcs/
10. Bureau of Health Informatics Office of Quality and Patient Safety NYS Department of Health. SPARCS Operations Guide (Version 1.2). 2016 Available from: https://www.health.ny.gov/statistics/sparcs/training/docs/sparcs_operations_guide.pdf

11. Arakaki L, Ngai S, Weiss D. Completeness of Neisseria meningitidis reporting in New York City, 1989–2010. Epidemiology and infection. 2016 Aug 1; 144(11):2374–81. https://doi.org/10.1017/S0950268816000406 PMID: 26984785

12. Bekelis K, Missios S, Coy S, MacKenzie TA. Scope of practice and outcomes of cerebrovascular procedures in children. Child's Nervous System. 2016 Nov 1; 32(11):2159–64. https://doi.org/10.1007/s00381-016-3114-2 PMID: 27193128

13. Bekelis K, Missios S, Coy S, MacKenzie TA. Comparison of outcomes of patients with inpatient or outpatient onset ischemic stroke. Journal of neurointerventional surgery. 2016 Jan 5:neurintsurg-2015.

14. Missios S, Bekelis K. Regional disparities in hospitalization charges for patients undergoing craniotomy for tumor resection in New York State: correlation with outcomes. Journal of neuro-oncology. 2016 Jun 1; 128(2):365–71. https://doi.org/10.1007/s11060-016-2122-0 PMID: 27072560

15. Kim H, Schwartz RM, Hirsch J, Silverman R, Liu B, Taioli E. Effect of Hurricane Sandy on Long Island emergency departments visits. Disaster medicine and public health preparedness. 2016 Jun 1; 10(03):344–50. https://doi.org/10.1017/dmp.2015.189 PMID: 26833178

16. He FT, De La Cruz NL, Olson D, Lim S, Seligson AL, Hall G, et al. 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 medicine and public health preparedness. 2016 Jun 1; 10(03):512–7. https://doi.org/10.1017/dmp.2016.89 PMID: 27292172

17. Chen X, Wang F. Integrative Spatial Data Analytics for Public Health Studies of New York State. In AMIA Annual Symposium Proceedings 2016 (Vol. 2016, p. 391). American Medical Informatics Association.

18. Chen X, Wang Y, Schoenfeld E, Saltz M, Saltz J, Wang F. Spatio-temporal analysis for New York State SPARCS data. In Proc. of 2017 AMIA Joint Summits on Translational Science. 2017 Mar27.

19. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. Available from: http://onlinelibrary.wiley.com/doi/10.3322/caac.21387/pdf

20. HCUP-US Tools & Software Page. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

21. American Cancer Society. Cancer facts & figures 2017. Atlanta: American Cancer Society. 2017.

22. U.S. Department of Health & Human Services. Summary of the HIPAA Privacy Rule. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

23. National Cancer Institute. Age. 2015. Available from: https://www.cancer.gov/about-cancer/causes-prevention/risk/age

24. Agrawal R, Srikant R. Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB. 1994 Sep 12 (Vol. 1215, pp. 487-499).

25. Python Software Foundation. apyori 1.1.1-Simple Apriori algorithm Implementation. Available from: https://pypi.python.org/pypi/apyori/1.1.1

26. Zaki MJ. Sequence mining in categorical domains: incorporating constraints. In Proceedings of the ninth international conference on Information and knowledge management 2000 Nov 6 (pp. 422-429). ACM.

27. Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Machine learning. 2001 Jan 1; 42(1):31–60. https://doi.org/10.1023/A:1007652502315

28. Welch HG, Chapko MK, James KE, Schwartz LM, Woloshin S. The role of patients and providers in the timing of follow-up visits. Journal of general internal medicine. 1999 Apr 1; 14(4):223–9. https://doi.org/10.1046/j.1525-1497.1999.00321.x PMID: 10203634

29. Buchta C, Hahsler M, Diaz D. arulesSequences: Mining Frequent Sequences. Available from: https://cran.r-project.org/web/packages/arulesSequences/index.html

30. SAS 9.4 Software. Available from: https://www.sas.com/en_us/software/sas9.html

31. Wang Y, Wang F. Association Rule Learning and Frequent Sequence Mining of Cancer Diagnoses in New York State. In Proceedings of the Third VLDB Workshop on Data Management and Analytics on Healthcare and Medicine (DMAH). 2017 Sep1.