Discovering Signature Disease Trajectories in Pancreatic Cancer and Soft-tissue Sarcoma from Longitudinal Patient Records

Liwei Wang[1†], Rui Li[1†], Andrew Wen[1], Qiuhao Lu[1], Jinlian Wang[1], Xiaoyang Ruan[1], Adriana Gamboa[2], Neha Malik[2], Christina L. Roland[2], Matthew H.G. Katz[2], Heather Lyu[2], Hongfang Liu[1*]

[1] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

[2] Department of Surgical Oncology, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

† Liwei Wang and Rui Li contributed equally.

* **Corresponding author:** Hongfang Liu, PhD

Postal address: 7000 Fannin Street Suite 600, Houston, TX 77030

Email: hongfang.liu@uth.tmc.edu

Telephone: 713-500-3900

Abstract

Understanding the disease trajectories of specific diseases can provide important clinical insights. In this paper, we aimed to discover signature disease trajectories of 3 rare cancer types: pancreatic cancer, soft tissue sarcoma (STS) of the trunk and extremity (STS-TE), and STS of the abdomen and retroperitoneum (STS-AR), leveraging IQVIA Oncology Electronic Medical Record. We identified significant diagnosis pairs in patients with these cancers through matched cohort sampling, statistical computation, right-tailed binomial hypothesis test, and visualized trajectories up to 3 progressions. Results included 266 significant diagnosis pairs for pancreatic cancer, 130 for STS-TE, and 118 for STS-AR. We further found 44 2-hop (i.e., 2-progression) and 136 3-hop trajectories before pancreatic cancer, 36 2-hop and 37 3-hop trajectories before STS-TE, and 17 2-hop and 5 3-hop trajectories before STS-AR. Meanwhile, we found 54 2-hop and 129 3-hop trajectories following pancreatic cancer, 11 2-hop and 17 3-hop trajectories following STS-TE, 5 2-hop and 0 3-hop trajectories following STS-AR. Systematic validation of discovered trajectories with the UTHealth Electronic Health Records confirmed the feasibility and reliability of our method. Our result suggested that some key clinical features can potentially serve as early markers of rare cancers. This approach is generalizable to other disease types and real-world longitudinal patient records.

INTRODUCTION

Rare cancers are defined as those with an incidence of fewer than 15 cases per 100,000 people per year.[1, 2] Pancreatic cancer and soft-tissue sarcoma (STS) are rare cancers that have an incidence of 8 cases per 100,000 people and 1.8 to 3 cases per 100,000 people, respectively.[3-5] Most clinicians have limited experience with these rare diseases, making diagnosis and treatment challenging. Large real-world data sources, such as electronic health records (EHRs), provide a massive amount of information that can potentially be leveraged to determine the patterns of diagnoses and treatments for rare tumors that can serve as clinical decision aids.

The study of large, diverse population-based datasets can help uncover disease trajectories.[6, 7] In one study, large-scale longitudinal registry data were used to identify 1,171 unique condensed disease trajectories that were grouped into 10 major clusters that represented the most common pathways for disease progression.[8] This approach has also been applied in the cancer domain; researchers have used a large patient registry to map the significant associations between diseases occurring before cancer diagnoses and created longitudinal disease trajectories that captured 168 temporal diagnosis patterns across 7 cancer types.[9] Additionally, an analysis of disease progression across more than 300 conditions in a dataset of 10.4 million inpatients uncovered an unexpected association between schizophrenia and rhabdomyolysis, highlighting the ability of these large-scale methods to reveal novel disease correlations.[10] The methods can also evaluate interconnected disease pathways; a recent study mapped temporal correlations of diseases commonly leading to hospital deaths.[11] These broad, overarching patterns of disease progression found in studies analyzing large-scale longitudinal records can help inform both research and clinical practice.[8-12]

Understanding general disease trajectories suggests the importance of identifying specific disease trajectories, particularly for rare diseases. Disease-specific trajectories can reveal subtle temporal patterns that can help clinicians refine diagnosis, treatment, and prognosis for a specific condition. These trajectories can also support the development of predictive models and algorithms to improve early detection and aid clinical decision-making. Nevertheless, studies on disease-specific trajectories remain limited. A recent study introduced a tool for visualizing and analyzing bladder cancer trajectories.[13] However, the selected study cohort with only bladder cancer patients allowed for commonly cooccurring medical problems to be filtered out, compromising the discovery of complete disease progression patterns. For example, hematuria, the most typical symptom of bladder cancer, was excluded from the final disease trajectory, which may not fully capture the clinical path leading to a bladder cancer diagnosis.[14,15]

To the best of our knowledge, there are no published studies on the signature disease trajectories of other rare cancers, including pancreatic cancer and STS. The aim of this study was to develop an approach to identify and describe the signature disease trajectories for pancreatic cancer, STS of the trunk and extremity (STS-TE), and STS of the abdomen and retroperitoneum (STS-AR) by leveraging large-scale EHR data.

METHODS

Data sources

IQVIA data were used to determine and visualize disease trajectories, and the UTHealth Observational Medical Outcomes Partnership (OMOP) common data model (CDM) dataset was used for trajectory validation. This study was approved by the Institutional Review Board of UTHealth.

IQVIA Oncology Electronic Medical Record

Data in the IQVIA Oncology Electronic Medical Record (EMR) dataset are sourced from oncology-specific EMR vendors at 3 types of sites: large practice, site services partner, and multiple practices/locations. Data from all sources are provided in a structured format and mapped to a common data model (CDM). These data include clinically rich information across various cancer types. The dataset includes clinical information from over 1.2 million patients from 2015 through 2023.

UTHealth OMOP CDM

EHR data from 2005 to 2024 from the McGovern Medical School at The University of Texas Health Sciences Center was transformed using the Observational Health Data Sciences and Informatics' OMOP CDM. Mapping to a CDM harmonizes data query formats across various data sources. Specifically, data from EHR legacy systems, including Allscripts Touchworks (pre-2021 for all clinical data) and GE Centricity (pre-2021 for billing) were combined with data from Epic (post-2021). These data were converted to the OMOP CDM on a nightly basis via an Apache Spark job. Structured data with available concept codes were directly converted from the source codes (International Disease Classification [ICD], Current Procedural Terminology [CPT], Healthcare Common Procedure Coding System [HCPCS], or Logical Observation Identifiers Names and Code [LOINC]) to the Athena concept vocabulary.[16] Currently, the UTHealth OMOP CDM dataset includes data on the conditions, measurements, procedures, drug exposures, and clinical narratives of 6.5 million patients.

Cohort construction

We constructed 3 cohorts of patients (pancreatic cancer and 2 types of STS) in both the IQVIA Oncology EMR and UTHealth OMOP CDM datasets. Because 60% of STS diagnoses occur in the extremities, 10% occur in the trunk, and 15% occur in the retroperitoneum, we created 2

STS cohorts, STS-TE and STS-AR, in each dataset.[14] For each disease, a list of ICD-10 diagnosis codes was first determined by clinical experts (Supplementary Data 1). In real-world datasets, ICD codes may not capture patients with a true STS diagnosis. To address this challenge, we further refined our strategies to build cohorts of actual STS cases. In the IQVIA dataset, we used both ICD-10 codes and sarcoma histology results to retrieve STS patients. In the UTHealth OMOP CDM dataset, we first retrieved potential STS patients using the ICD-10 codes and then confirmed that the keyword "sarcoma" appeared in the patient's clinical notes.

Computation strategy

Our computation strategy included significant diagnosis pair mining and disease trajectory visualization. Figure 1 shows an overview of the computation strategy.

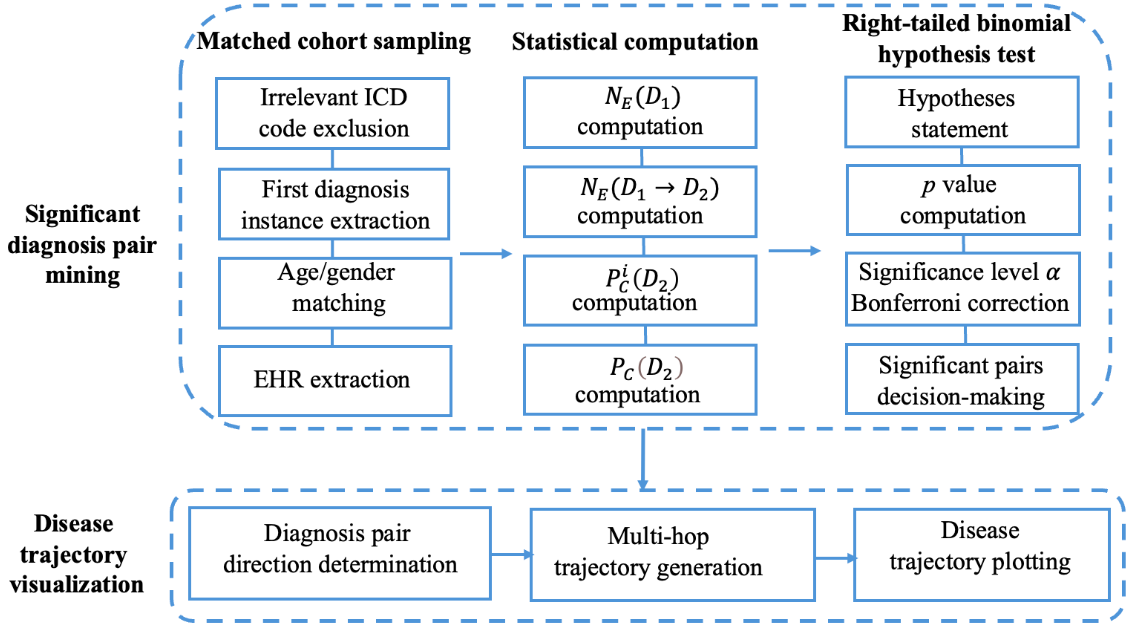
Significant diagnosis pair mining
Matched cohort sampling
Irrelevant ICD code exclusion
First diagnosis instance extraction
Age/gender matching
EHR extraction
Statistical computation
$N_E(D_1)$ computation
$N_E(D_1 \rightarrow D_2)$ computation
$P_C^i(D_2)$ computation
$P_C(D_2)$ computation
Right-tailed binomial hypothesis test
Hypotheses statement
p value computation
Significance level α Bonferroni correction
Significant pairs decision-making
Disease trajectory visualization
Diagnosis pair direction determination
Multi-hop trajectory generation
Disease trajectory plotting

Figure 1. Overview of the computation strategy. EHR, electronic health record; ICD, International Classification of Diseases.

Significant diagnosis pair mining

In identifying significant diagnosis pairs, our primary goal was to find the pair (D_1, D_2), in which the conditional probability $P_e(D_2|D_1)$ in patients with the cancer type (i.e., pancreatic cancer, STS-TE, or STS-AE) was higher than the marginal probability $P_c(D_2)$ in the general population. This was achieved through matched cohort sampling, probability computation, and a right-tailed binomial hypothesis test.

Matched cohort sampling

To speed up computation, we extracted the first 4 digits of the full ICD-10 codes and excluded the irrelevant ICD-10 codes starting with “O__”, “P__”, “R__”, “S__”, “T__”, “V__”, “W__”, “X__”, “Y__”, and “Z__”. We retained only the first diagnosis of for any disease occurred in the cancer cohort and obtained all possible diagnosis pairs. Given a diagnosis pair (D_1, D_2), we defined the exposed group (E) and the comparison group (C). The exposed group contained patients diagnosed with both the cancer type (i.e., pancreatic cancer, STS-TE, or STS-AE) and D_1. We extracted all diagnosis records starting from the diagnosis of D_1 through the next 3 years in order to calculate $P_E(D_2|D_1)$. Each patient in the exposed group was matched to individuals from the entire IQVIA dataset according to age and gender to create a comparison group. We then extracted diagnosis records for these matched individuals within 3 years of their matching point. For example, we matched a female who was diagnosed with D1 at 60 in the exposed group to the females who were diagnosed with any disease at 60. We then extracted all diagnosis records of the matched patients within 3 years in the IQVIA. We pre-calculate the temporal diagnosis pair numbers across all 3 cancer types, and the average time interval for each pair. We found that there were 607,366 temporal diagnosis pairs in total and 87.3% of pairs occurred within a 3-year period. This justifies the use of the 3-year time window.

Notably, the patients in the exposed group had to be diagnosed with D_1 but did not need to have a diagnosis of D_2, as the aim of collecting data from the exposed group was to calculate the probability $P_E(D_2|D_1)$ to compare with the marginal probability $P_C(D_2)$. Patients in the comparison group did not have to have a diagnosis of D_1 or D_2.

Statistical computation

Given the diagnosis pair (D_1, D_2) and the corresponding exposed group (E) and comparison group (C), $N_E(D_1)$ represents the number of patients who were diagnosed with D_1 in the exposed group, and $N_E(D_1 \rightarrow D_2)$ represents the number of patients who were diagnosed with D_1 and diagnosed with D_2 within 3 years. $P_C(D_2)$ represents the probability of a patient in the comparison group being diagnosed with D_2 (Figure 2).

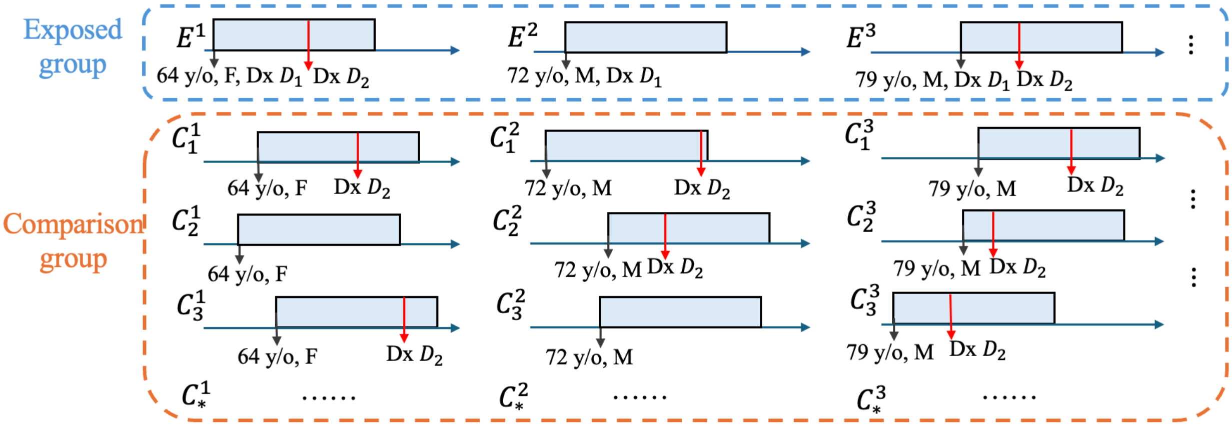

Figure 2. Matched cohort sampling and probability computation. The blue dashed box represents the exposed group, consisting of 3 example patients (E^1, E^2, E^3). All of these patients were diagnosed with D_1 and may or may not have been diagnosed with D_2. The orange box represents the comparison group, containing multiple patients who were age- and gender-

matched to the patients in the exposed group (C_*^1, C_*^2, C_*^3). It is important to note that although only 3 matched examples displayed, the comparison group includes all patients meeting the matching criteria from the entire IQVIA dataset. The superscript $*$ denotes the index of the patient in the exposed group. F, female; M, male; y/o, years old; Dx, diagnosis

For each patient in the exposed group, E^i, $P_C^i(D_2)$ represents the probability of being diagnosed with D_2 across all patients in that exposed patient's comparison group. The overall probability $P_c(D_2)$ is computed as the average of $P_C^i(D_2)$ across all patients in the exposed group. If fewer than 2,000 patients were matched for any given E^i, the match was deemed unrepresentative and the E^i was excluded from the exposed group when statistically analyzing the (D_1, D_2) pair.

To optimize the computation process, we performed a pre-filtering step, keeping only the diagnosis pairs that were present in more than 0.75% of patients with the cancer of interest. This percentage could be adjusted based on computational capacity

Right-tailed binomial hypothesis test

To determine whether the pair (D_1, D_2) was significant, we conducted a right-tailed binomial hypothesis test. The hypotheses were:

$$\begin{cases} H_0: P_E(D_2|D_1) \ \leq P_C(D_2) \\ H_1: P_E(D_2|D_1) > P_C(D_2) \end{cases}$$

We used a binomial distribution to model $N_E(D_1 \rightarrow D_2)$. In this case, the number of trials is $N_E(D_1)$, the success probability under the null hypothesis is $P_C(D_2)$, and the observed number of successes is $N_E(D_1 \rightarrow D_2)$. The *p* value is the observed probability when $N_E(D_1 \rightarrow D_2)$ is greater than $P_C(D_2) * N_E(D_1)$. We first set the significance level α to 0.05. Because there were thousands of diagnosis pairs, we performed the Bonferroni correction[17] to adjust the significance level to control the overall type I error rate across multiple comparisons. The adjusted significance level α' was computed, with $\alpha' = \alpha/m$, where m is the number of diagnosis pairs. Then, we computed $p = P(N_E(D_1 \to D_2) > P_C(D_2) * N_E(D_1))$. If $p < \alpha'$, H_0 was rejected, i.e., if $N_E(D_1 \to D_2)$ was significantly greater than $P_C(D_2) * N_E(D_1)$ and thus $P_E(D_2|D_1) > P_C(D_2)$, the pair (D_1, D_2) was significant.

Disease trajectory visualization

Disease trajectory visualization was achieved through the determination of the directionality of the diagnosis pairs, multi-hop trajectory generation, and disease trajectory plotting. A “hop” refers to a progression from one disease to another (i.e., a new diagnosis).

Determination of the directionality of diagnosis pairs

The diagnosis pairs (D_1, D_2) and (D_2, D_1) both could have been identified as significant. In this case, we needed to determine the overall directionality of that diagnosis pair. $N_E(D_1 \to D_2)$ denotes the number of patients diagnosed with D_1 and then diagnosed with D_2, and $N_E(D_2 \to D_1)$ denotes the number of patients diagnosed with D_2 and then diagnosed with D_1. We set $N_E(D_1, D_2) = N_E(D_1 \to D_2) + N_E(D_2 \to D_1)$. Then, we compared $N_E(D_1 \to D_2)$ and $N_E(D_2 \to D_1)$ with $0.5 * N_E(D_1, D_2)$, and performed the right-tailed binomial hypothesis test to determine the overall directionality of the disease trajectory. If the directionality from $D_1 \to D_2$ was significant, it indicated that the diagnosis of D_1 preceded that of D_2, establishing a directional link from D_1 to D_2 in the disease trajectory. If the directionality from $D_2 \to D_1$ was significant, it indicated that the diagnosis of D_2 preceded that of D_1, establishing a directional link from D_2 to D_1 in the disease trajectory. If neither directionality was significant, it suggested that D_1 and D_2 were co-occurring.

Multi-hop trajectory generation

Long trajectories with multiple hops were constructed by sequentially concatenating 1-hop (diagnosis pair) trajectories with directionalities. For instance, if both $D_1 \rightarrow D_2$ and $D_2 \rightarrow D_3$ were significant, $D_1 \rightarrow D_2 \rightarrow D_3$ formed a 2-hop trajectory with 3 nodes. If this trajectory was included in the EHR data of patients with the disease of interest, we classified $D_1 \rightarrow D_2 \rightarrow D_3$ as a valid 2-hop trajectory. Once all 2-hop trajectories were identified, we concatenated them with the 1-hop trajectories to generate potential 3-hop trajectories. These 3-hop trajectories were then filtered so that only those present in at least 1 patient's EHR data were retained. We continued this process iteratively, expanding the trajectory length until no additional trajectories were identified.

Disease trajectory plotting

After obtaining the diagnosis pairs with and without directionality, we used Gephi 0.10,[18] an open-source software platform designed for visualizing and analyzing large networks and graphs, to visualize the disease trajectories.

Trajectory validation

We systematically validated trajectories with 1 hop, 2 hops, and 3 hops with and without directionality using patient data encompassing all 3 types of cancer from the UTHealth OMOP CDM. Specifically, we kept only the first diagnosis date of any disease occurred in all patients with a pancreatic cancer, STS-TE, or STS-AR diagnosis. We then compared the temporally ordered diagnoses of cancer patients with the discovered trajectories, and the overlapping trajectories were regarded as validated. For example, in the 2-hop trajectory of unspecified anemia (D64.9), unspecified iron-deficiency anemia (D50.9), and STS-TE, if a patient had temporally ordered diagnoses of these diseases, then the trajectory was validated.

RESULTS

Patient characteristics

Table 1 shows the demographic characteristics of the cancer cohorts in the IQVIA and UTHealth OMOP CDM. The IQVIA cohort had approximately 5 times more patients than the UTHealth OMOP CDM cohort.

Table 1. Demographic characteristics of patients by cancer type and dataset (IQVIA vs. UTHealth OMOP CDM).

	Pancreatic cancer		**STS-TE**		**STS-AR**	
Characteristic	**IQVIA (n = 13,606)**	**UTHealth (n = 2,568)**	**IQVIA (n = 4,536)**	**UTHealth (n = 860)**	**IQVIA (n = 1,003)**	**UTHealth (n =179)**
Gender						
Female	7,171 (52.7)	1,241 (48.3)	2,194 (48.4)	539 (62.7)	497 (49.6)	96 (53.6)
Male	6,426 (47.2)	1,326 (51.6)	2,342 (51.6)	321 (37.3)	504 (50.2)	83 (46.4)
Missing	9 (0.1)	1 (0.1)	0 (0)	0 (0)	2 (0.2)	0 (0)
Average age at 1st diagnosis, years	68	64	58	55	61	55
0–10	1 (0.01)	2 (0.1)	62 (1.4)	16 (1.9)	6 (0.6)	6 (3.4)
10–20	9 (0.1)	6 (0.2)	113 (2.5)	24 (2.8)	10 (1.0)	6 (3.4)
20–30	27 (0.2)	24 (0.9)	221 (4.9)	61 (7.1)	32 (3.2)	12 (6.7)
30–40	143 (1.1)	70 (2.7)	331 (7.3)	72 (8.4)	48 (4.8)	10 (5.6)

40–50	554 (4.1)	195 (7.6)	528 (11.6)	115 (13.4)	96 (9.6)	22 (12.3)
50–60	2,026 (14.9)	541 (21.1)	803 (17.7)	137 (15.9)	175 (17.4)	30 (16.7)
60–70	4,224 (31.0)	806 (31.4)	1041 (22.9)	203 (23.6)	293 (29.2)	50 (27.9)
70–80	4,864 (35.7)	643 (25.0)	1080 (23.8)	167 (19.4)	273 (27.2)	31 (17.3)
80–90	1,758 (12.9)	251 (9.8)	357 (7.9)	57 (6.6)	70 (7.0)	12 (6.7)
Race						
African American	1,145 (8.4)	548 (21.3)	464 (10.2)	181 (21.0)	117 (11.7)	32 (17.9)
Asian	169 (1.2)	102 (4.9)	74 (1.6)	27 (3.1)	24 (2.4)	9 (5.0)
White	8,486 (62.4)	1,052 (41.0)	3,069 (67.7)	403 (46.9)	639 (63.7)	81 (45.3)
Others	538 (4.0)	5 (0.3)	144 (3.2)	0 (0)	49 (4.9)	0 (0)
Unknown	3,268 (24.0)	861 (33.5)	785 (17.3)	249 (29.0)	174 (17.3)	57 (31.8)

STS-AR, soft-tissue sarcoma of the abdomen and retroperitoneum; STS-TE, soft-tissue sarcoma of the trunk and extremity; UTHealth, University of Texas Health Science Center. All values are n (%) unless otherwise indicated. Others include American Indians or Alaska Natives, Native Hawaiian, or Other Pacific Islanders.

Significant diagnosis pairs

Table 2 shows the relevant values for calculating significant diagnosis pairs for the 3 cancer categories. Among the 13,606 pancreatic cancer patients, 389,919 ICD-10 diagnosis pairs were initially found. After pre-filtering, only 469 diagnosis pairs remained. The Bonferroni-corrected significance level was set to 1.28×10^{-7} with $\alpha' = \alpha/N$. We then conducted a right-tailed binomial hypothesis test, identifying 266 diagnosis pairs as significant. Regarding directionality, 16 diagnosis pairs were found to have no directionality (i.e., were co-occurring), and 250 diagnosis pairs were found to be directional. Among those 250 pairs, 39 pairs followed pancreatic cancer, 46 pairs preceded pancreatic cancer, and 165 pairs contained no pancreatic cancer diagnoses. There were 130 significant pairs, with 124 being directional for STS-TE, and 118 significant pairs, with 113 being directional for STS-AR. Tables S1, S6 and S11 of the Supplementary Data 2 contain the significant pairs for each type of cancer.

Table 2. Calculation of diagnosis pairs.

				No. of statistically significant pairs			
					Directionality		
Cancer type	**Before prefiltering**	**After prefiltering**	α'	**No directionality**	**Pointing to the cancer**	**Going from the cancer**	**With no cancer**
Pancreatic	389,919	469	1.28×10^{-7}	16	46	39	165
STS-TE	248,304	383	2.01×10^{-7}	6	62	13	49
STS-AR	55,766	411	8.97×10^{-7}	5	59	5	49

STS-AR, soft-tissue sarcoma of the abdomen and retroperitoneum; STS-TE, soft-tissue sarcoma of the trunk and extremity.

Disease trajectories

Figure 3, Figure 4, and Figure 5 show the significant disease trajectories of pancreatic cancer, STS-TE, and STS-AR, respectively. In terms of diagnosis pairs with the highest edge strength (i.e., most common trajectories), the top 3 diagnoses that preceded pancreatic cancer were primary hypertension (I10) with 954 patients accounting for 7% of all pancreatic patients in IQVIA, hyperlipidemia (E78.5) (595, 4.4%), and type 2 diabetes mellitus (E11.9) (553, 4.1%). The top 3 diagnoses that followed pancreatic cancer were dehydration (E86.0) (1828, 13.4%), neoplasm-related pain (G89.3) (1301, 9.6%), and Secondary malignant neoplasm of liver and intrahepatic bile ducts (C78.7) (1197, 8.8%). A total of 43 disease diagnoses preceded both STS-TE and STS-AR, accounting for 69.4% and 72.9% of prior diagnoses, respectively. The top 3 diagnoses that preceded STS-TE were primary hypertension (I10) (285, 6.3%); pain in a joint (M25.5) (219, 4.8%); and pain in a limb, hand, foot, finger, or toe (M79.6) (175, 3.9%). The top 3 diagnoses that preceded STS-AR were primary hypertension (I10) (69, 6.9%), unspecified anemia (D64.9) (42, 4.2%), and pain in a joint (M25.5) (41, 4.1%). A total of 2 disease diagnoses followed both STS-TE and STS-AR, accounting for 15.4% and 40.0% of prior diagnoses, respectively. The top 3 diagnoses that followed STS-TE were secondary malignant neoplasm of the lung (C78.0) (238, 5.2%), neoplasm-related pain (G89.3) (229, 5%), and agranulocytosis secondary to cancer chemotherapy (D70.1) (162, 3.6%). The top 3 diagnoses that followed STS-AR were malignant neoplasm of connective and soft tissue (C49.9) (101, 10%), unspecified neoplasm-related pain (G89.3) (58, 5.8%), and secondary malignant neoplasm of the lung (C78.0) (38, 3.8%).

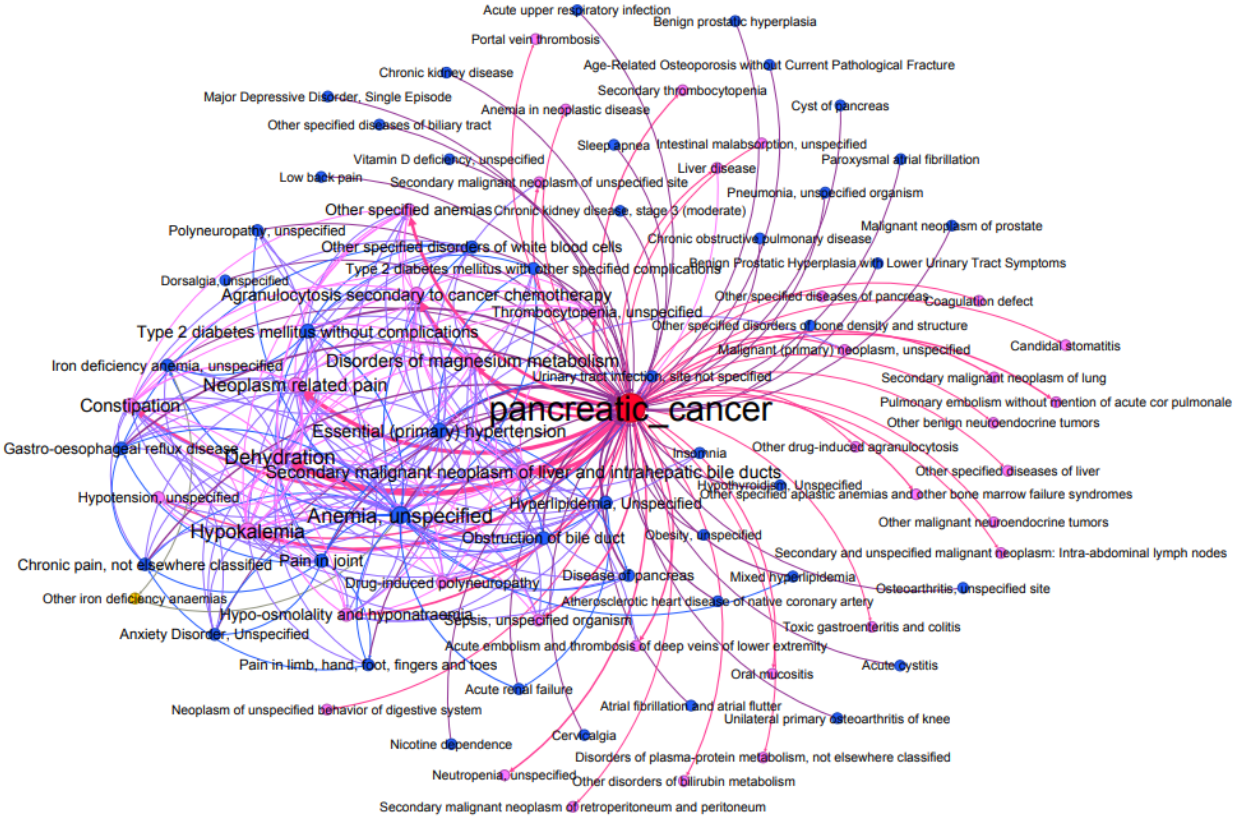
Acute upper respiratory infection
Benign prostatic hyperplasia
Portal vein thrombosis
Chronic kidney disease
Age-Related Osteoporosis without Current Pathological Fracture
Major Depressive Disorder, Single Episode
Secondary thrombocytopenia
Anemia in neoplastic disease
Cyst of pancreas
Other specified diseases of biliary tract
Sleep apnea
Intestinal malabsorption, unspecified
Vitamin D deficiency, unspecified
Paroxysmal atrial fibrillation
Liver disease
Low back pain
Secondary malignant neoplasm of unspecified site
Pneumonia, unspecified organism
Other specified anemias
Chronic kidney disease, stage 3 (moderate)
Polyneuropathy, unspecified
Malignant neoplasm of prostate
Chronic obstructive pulmonary disease
Other specified disorders of white blood cells
Type 2 diabetes mellitus with other specified complications
Benign Prostatic Hyperplasia with Lower Urinary Tract Symptoms
Dorsalgia, unspecified
Agranulocytosis secondary to cancer chemotherapy
Other specified diseases of pancreas
Coagulation defect
Thrombocytopenia, unspecified
Type 2 diabetes mellitus without complications
Other specified disorders of bone density and structure
Candidal stomatitis
Malignant (primary) neoplasm, unspecified
Iron deficiency anemia, unspecified
Disorders of magnesium metabolism
Urinary tract infection, site not specified
Secondary malignant neoplasm of lung
Neoplasm related pain
Constipation
pancreatic_cancer
Pulmonary embolism without mention of acute cor pulmonale
Other benign neuroendocrine tumors
Essential (primary) hypertension
Gastro-oesophageal reflux disease
Insomnia
Other drug-induced agranulocytosis
Dehydration
Other specified diseases of liver
Secondary malignant neoplasm of liver and intrahepatic bile ducts
Hypothyroidism, Unspecified
Other specified aplastic anemias and other bone marrow failure syndromes
Hypotension, unspecified
Hyperlipidemia, Unspecified
Anemia, unspecified
Other malignant neuroendocrine tumors
Hypokalemia
Obstruction of bile duct
Obesity, unspecified
Secondary and unspecified malignant neoplasm: Intra-abdominal lymph nodes
Chronic pain, not elsewhere classified
Pain in joint
Disease of pancreas
Mixed hyperlipidemia
Osteoarthritis, unspecified site
Drug-induced polyneuropathy
Other iron deficiency anaemias
Atherosclerotic heart disease of native coronary artery
Hypo-osmolality and hyponatraemia
Sepsis, unspecified organism
Toxic gastroenteritis and colitis
Anxiety Disorder, Unspecified
Acute embolism and thrombosis of deep veins of lower extremity
Pain in limb, hand, foot, fingers and toes
Acute cystitis
Oral mucositis
Acute renal failure
Atrial fibrillation and atrial flutter
Neoplasm of unspecified behavior of digestive system
Unilateral primary osteoarthritis of knee
Cervicalgia
Nicotine dependence
Disorders of plasma-protein metabolism, not elsewhere classified
Neutropenia, unspecified
Other disorders of bilirubin metabolism
Secondary malignant neoplasm of retroperitoneum and peritoneum

Figure 3. Disease trajectories of patients with pancreatic cancer. The thickness of each trajectory (i.e., line) represents the number of patients associated with that trajectory; thicker arrows represent trajectories containing more patients. The nodes in the trajectories are categorized as follows: red represents pancreatic cancer; blue represents nodes that point to pancreatic cancer, indicating that the diagnosis nodes precede the pancreatic cancer diagnosis; pink refers to nodes pointed away from pancreatic cancer, indicating that the diagnosis occurred after the pancreatic cancer diagnosis; and yellow indicates that a node was not linked to pancreatic cancer.

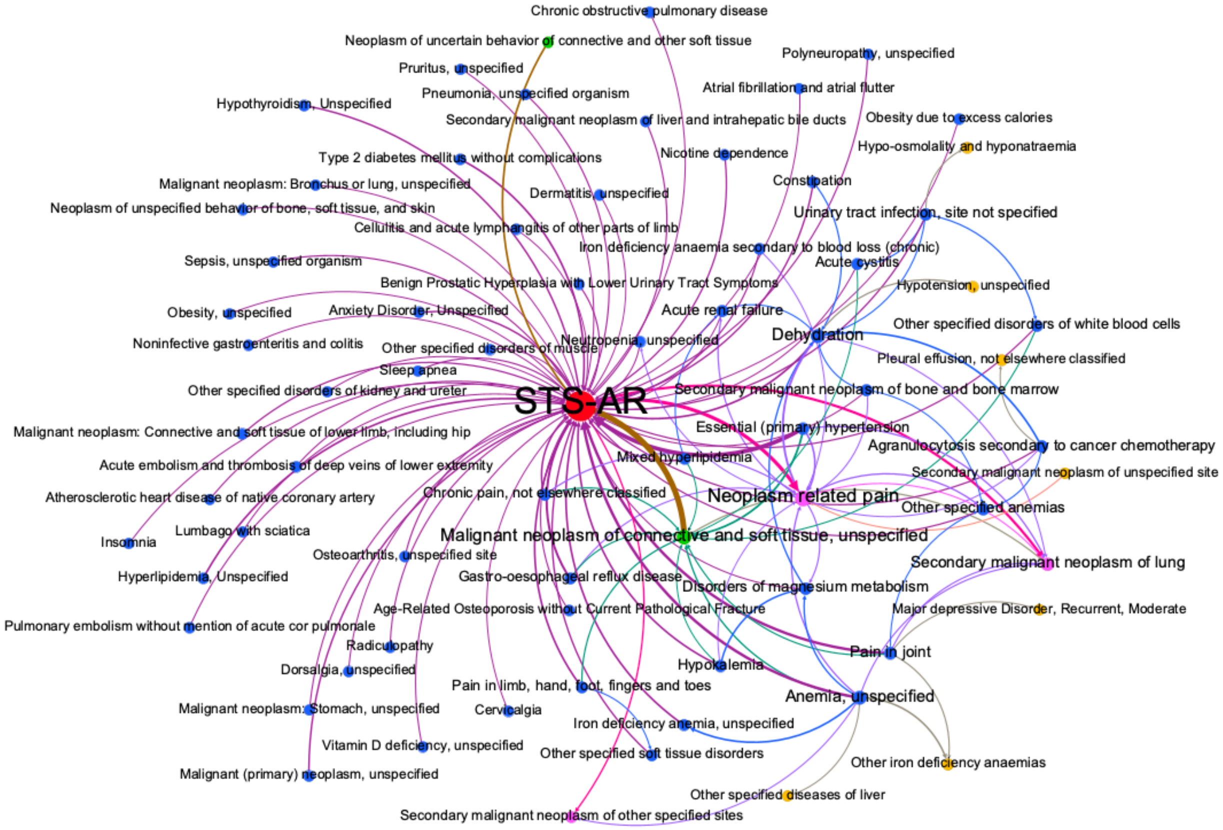
Chronic obstructive pulmonary disease
Neoplasm of uncertain behavior of connective and other soft tissue
Polyneuropathy, unspecified
Pruritus, unspecified
Atrial fibrillation and atrial flutter
Pneumonia, unspecified organism
Hypothyroidism, Unspecified
Secondary malignant neoplasm of liver and intrahepatic bile ducts
Obesity due to excess calories
Hypo-osmolality and hyponatraemia
Nicotine dependence
Type 2 diabetes mellitus without complications
Constipation
Malignant neoplasm: Bronchus or lung, unspecified
Dermatitis, unspecified
Neoplasm of unspecified behavior of bone, soft tissue, and skin
Urinary tract infection, site not specified
Cellulitis and acute lymphangitis of other parts of limb
Iron deficiency anaemia secondary to blood loss (chronic)
Acute cystitis
Sepsis, unspecified organism
Benign Prostatic Hyperplasia with Lower Urinary Tract Symptoms
Hypotension, unspecified
Obesity, unspecified
Anxiety Disorder, Unspecified
Acute renal failure
Other specified disorders of white blood cells
Dehydration
Noninfective gastroenteritis and colitis
Other specified disorders of muscle
Neutropenia, unspecified
Pleural effusion, not elsewhere classified
Sleep apnea
Other specified disorders of kidney and ureter
STS-AR
Secondary malignant neoplasm of bone and bone marrow
Malignant neoplasm: Connective and soft tissue of lower limb, including hip
Essential (primary) hypertension
Agranulocytosis secondary to cancer chemotherapy
Acute embolism and thrombosis of deep veins of lower extremity
Mixed hyperlipidemia
Secondary malignant neoplasm of unspecified site
Atherosclerotic heart disease of native coronary artery
Chronic pain, not elsewhere classified
Neoplasm related pain
Other specified anemias
Lumbago with sciatica
Malignant neoplasm of connective and soft tissue, unspecified
Insomnia
Osteoarthritis, unspecified site
Secondary malignant neoplasm of lung
Hyperlipidemia, Unspecified
Gastro-oesophageal reflux disease
Disorders of magnesium metabolism
Age-Related Osteoporosis without Current Pathological Fracture
Major depressive Disorder, Recurrent, Moderate
Pulmonary embolism without mention of acute cor pulmonale
Radiculopathy
Pain in joint
Dorsalgia, unspecified
Hypokalemia
Pain in limb, hand, foot, fingers and toes
Anemia, unspecified
Malignant neoplasm: Stomach, unspecified
Cervicalgia
Iron deficiency anemia, unspecified
Vitamin D deficiency, unspecified
Other specified soft tissue disorders
Other iron deficiency anaemias
Malignant (primary) neoplasm, unspecified
Other specified diseases of liver
Secondary malignant neoplasm of other specified sites

Figure 4. Disease trajectories of soft-tissue sarcoma of the abdomen and retroperitoneum (STS-AR). The thickness of each trajectory (i.e., line) represents the number of patients associated with that trajectory; thicker arrows represent trajectories containing more patients. The disease nodes in the trajectories are categorized as follows: red represents STS-AR; blue represents nodes that point to STS-AR, indicating that the diagnosis nodes precede the STS-AR diagnosis; pink refers to nodes pointed away from STS-AR, indicating that the diagnosis occurred after the STS-AR diagnosis; green represents nodes linked to STS-AR with no directionality, indicating that the disease co-occurred with STS-AR; and yellow indicates that a node was not linked to STS-AR.

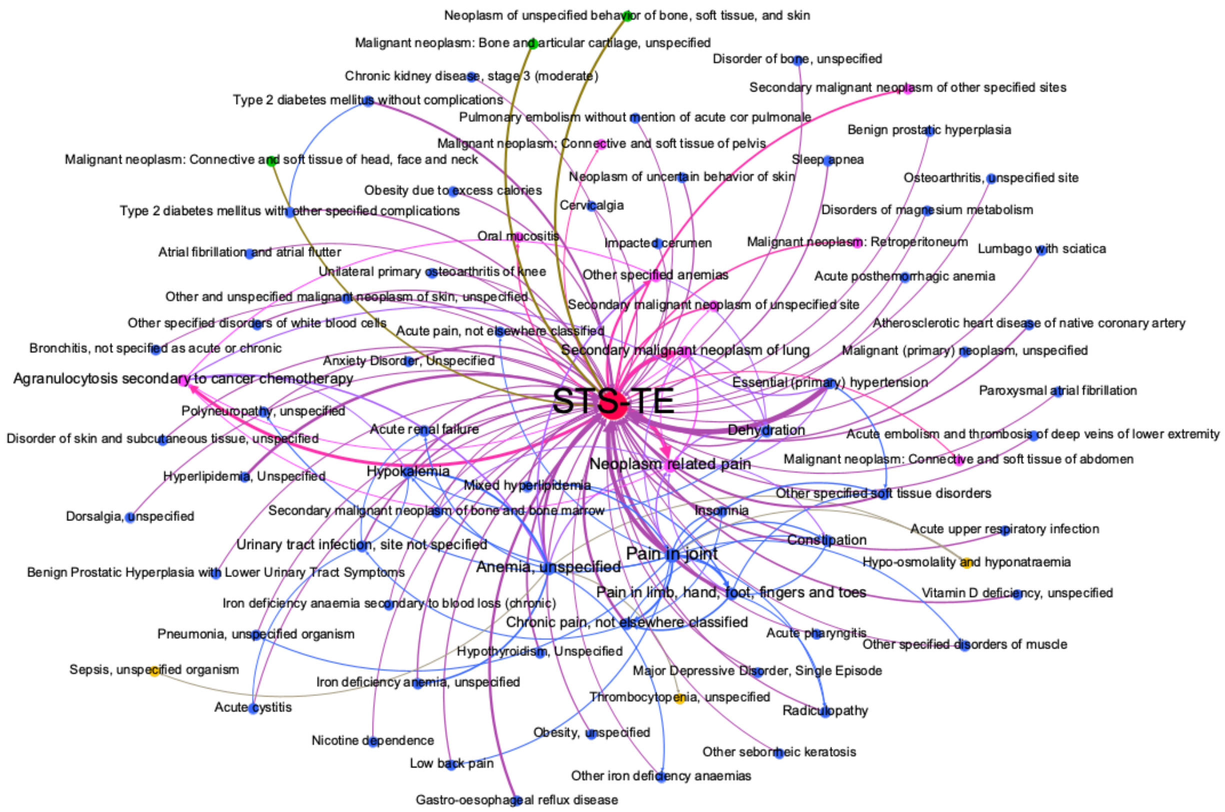
Neoplasm of unspecified behavior of bone, soft tissue, and skin
Malignant neoplasm: Bone and articular cartilage, unspecified
Disorder of bone, unspecified
Chronic kidney disease, stage 3 (moderate)
Secondary malignant neoplasm of other specified sites
Type 2 diabetes mellitus without complications
Pulmonary embolism without mention of acute cor pulmonale
Benign prostatic hyperplasia
Malignant neoplasm: Connective and soft tissue of pelvis
Malignant neoplasm: Connective and soft tissue of head, face and neck
Sleep apnea
Neoplasm of uncertain behavior of skin
Osteoarthritis, unspecified site
Obesity due to excess calories
Cervicalgia
Disorders of magnesium metabolism
Type 2 diabetes mellitus with other specified complications
Oral mucositis
Impacted cerumen
Malignant neoplasm: Retroperitoneum
Lumbago with sciatica
Atrial fibrillation and atrial flutter
Unilateral primary osteoarthritis of knee
Other specified anemias
Acute posthemorrhagic anemia
Other and unspecified malignant neoplasm of skin, unspecified
Secondary malignant neoplasm of unspecified site
Other specified disorders of white blood cells
Acute pain, not elsewhere classified
Atherosclerotic heart disease of native coronary artery
Bronchitis, not specified as acute or chronic
Secondary malignant neoplasm of lung
Malignant (primary) neoplasm, unspecified
Anxiety Disorder, Unspecified
Agranulocytosis secondary to cancer chemotherapy
STS-TE
Essential (primary) hypertension
Paroxysmal atrial fibrillation
Polyneuropathy, unspecified
Acute renal failure
Dehydration
Acute embolism and thrombosis of deep veins of lower extremity
Disorder of skin and subcutaneous tissue, unspecified
Neoplasm related pain
Malignant neoplasm: Connective and soft tissue of abdomen
Hypokalemia
Hyperlipidemia, Unspecified
Mixed hyperlipidemia
Other specified soft tissue disorders
Secondary malignant neoplasm of bone and bone marrow
Insomnia
Dorsalgia, unspecified
Acute upper respiratory infection
Constipation
Urinary tract infection, site not specified
Pain in joint
Hypo-osmolality and hyponatraemia
Anemia, unspecified
Benign Prostatic Hyperplasia with Lower Urinary Tract Symptoms
Pain in limb, hand, foot, fingers and toes
Vitamin D deficiency, unspecified
Iron deficiency anaemia secondary to blood loss (chronic)
Chronic pain, not elsewhere classified
Acute pharyngitis
Pneumonia, unspecified organism
Other specified disorders of muscle
Hypothyroidism, Unspecified
Major Depressive Disorder, Single Episode
Sepsis, unspecified organism
Iron deficiency anemia, unspecified
Thrombocytopenia, unspecified
Acute cystitis
Radiculopathy
Obesity, unspecified
Nicotine dependence
Other seborrheic keratosis
Low back pain
Other iron deficiency anaemias
Gastro-oesophageal reflux disease

Figure 5. Disease trajectories of soft-tissue sarcoma of the trunk and extremity (STS-TE). The thickness of each trajectory (i.e., line) represents the number of patients associated with that trajectory; thicker arrows represent trajectories containing more patients. The disease nodes in the trajectories are categorized as follows: red represents STS-TE; blue represents nodes that point to STS-TE, indicating that the diagnosis nodes precede the STS-TE diagnosis; pink refers to nodes pointed away from STS-TE, indicating that the diagnosis occurred after the STS-TE diagnosis; green represents nodes linked to STS-AR with no directionality, indicating that the disease co-occurred with STS-TE; and yellow indicates that a node was not linked to STS-TE.

Table 3 shows the number of long trajectories for each cancer type. Figure 6 shows the example disease trajectories that preceded a pancreatic cancer, STS-TE, or STS-AR diagnosis found in most patients. For example, type 2 diabetes, gastroesophageal reflux disease and anemia occurred before pancreatic cancer. Hyperlipidemia occurred before STS-TE. Figure 7 shows the example disease trajectories following pancreatic cancer, STS-TE, or STS-AR diagnosis found in most patients. Supplementary Data 2 (Tables S2-S14 excluding S6 and S11) contains the detailed data of the long trajectories (i.e., those with multiple hops) discovered from the IQVIA Oncology EMR for pancreatic cancer, STS-AR, and STS-TE. For example, hypertension occurred before STS.

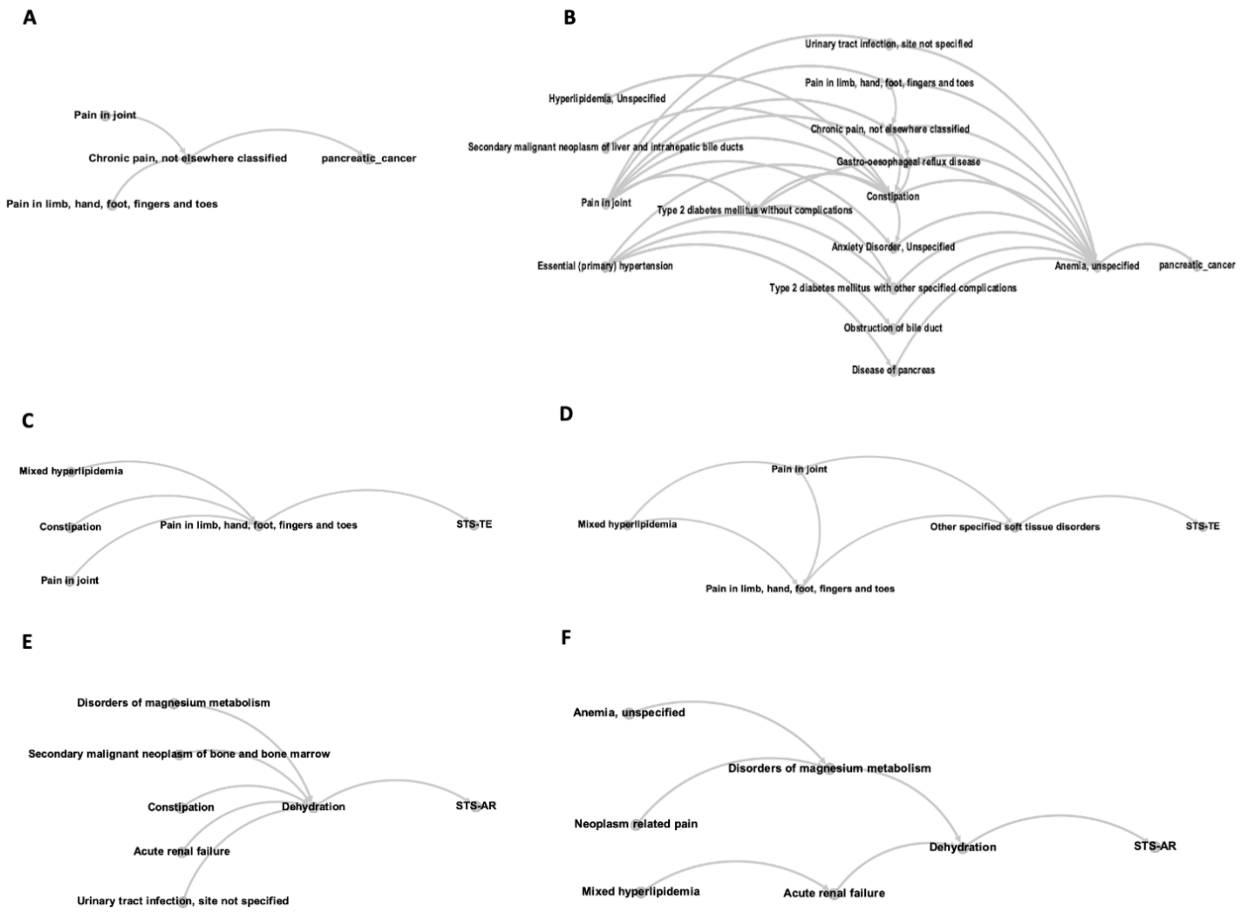
A
Pain in joint
Chronic pain, not elsewhere classified
pancreatic_cancer
Pain in limb, hand, foot, fingers and toes
B
Urinary tract infection, site not specified
Pain in limb, hand, foot, fingers and toes
Hyperlipidemia, Unspecified
Chronic pain, not elsewhere classified
Secondary malignant neoplasm of liver and intrahepatic bile ducts
Gastro-oesophageal reflux disease
Pain in joint
Type 2 diabetes mellitus without complications
Constipation
Anxiety Disorder, Unspecified
Essential (primary) hypertension
Anemia, unspecified
pancreatic_cancer
Type 2 diabetes mellitus with other specified complications
Obstruction of bile duct
Disease of pancreas
C
Mixed hyperlipidemia
Constipation
Pain in limb, hand, foot, fingers and toes
STS-TE
Pain in joint
D
Pain in joint
Mixed hyperlipidemia
Other specified soft tissue disorders
STS-TE
Pain in limb, hand, foot, fingers and toes
E
Disorders of magnesium metabolism
Secondary malignant neoplasm of bone and bone marrow
Constipation
Dehydration
STS-AR
Acute renal failure
Urinary tract infection, site not specified
F
Anemia, unspecified
Disorders of magnesium metabolism
Neoplasm related pain
Dehydration
STS-AR
Mixed hyperlipidemia
Acute renal failure

Figure 6. Possible disease trajectories before pancreatic cancer, STS-TE and STS-AR diagnoses. A) Trajectories with 2 hops before a pancreatic cancer diagnosis. B) Trajectories

with 3 hops before a pancreatic cancer diagnosis. C) Trajectories with 2 hops before an STS-TE diagnosis. D) Trajectories with 3 hops before an STS-TE diagnosis. E) Trajectories with 2 hops before an STS-AR diagnosis. F) Trajectories with 3 hops before an STS-AR diagnosis.

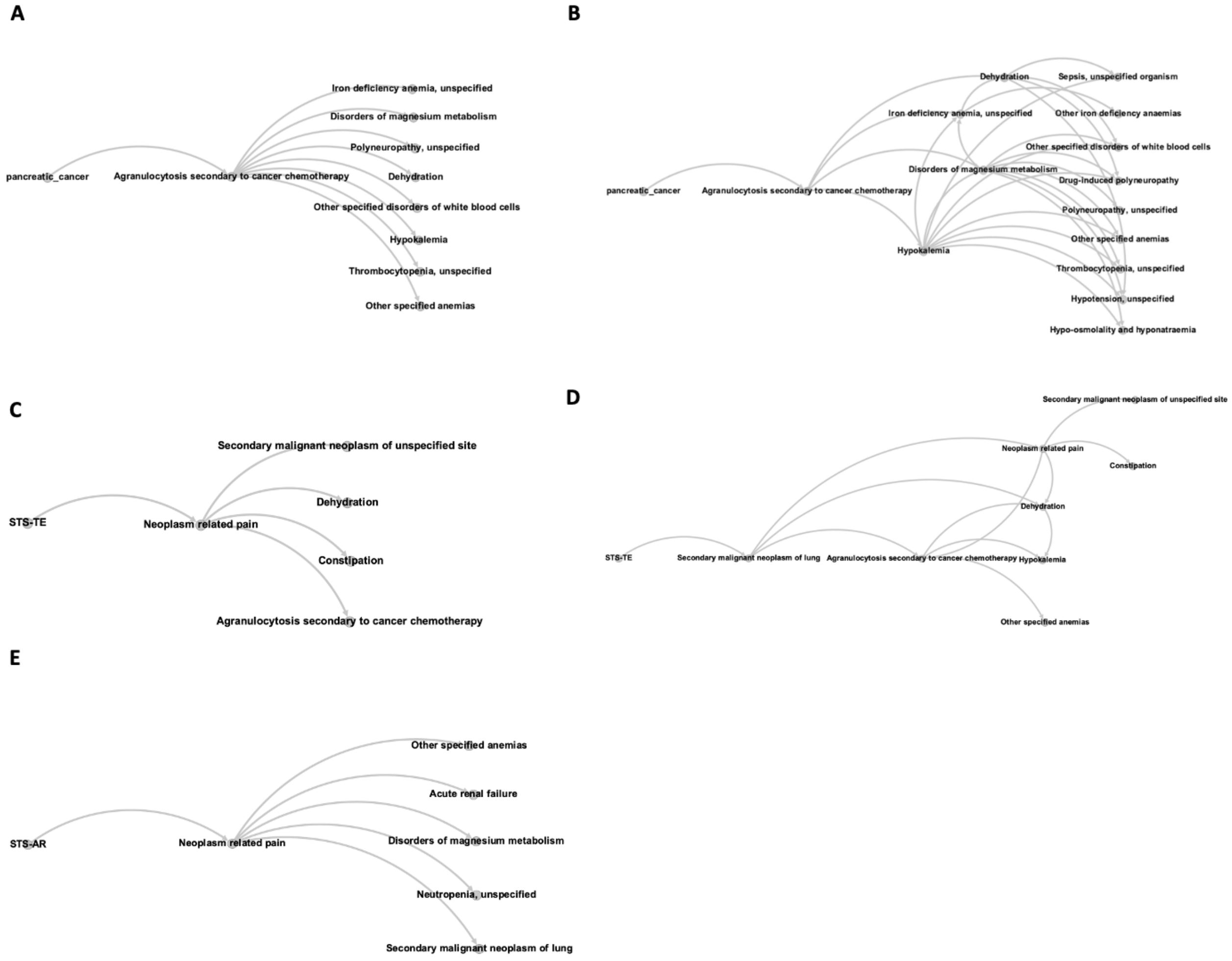
A
pancreatic_cancer
Agranulocytosis secondary to cancer chemotherapy
Iron deficiency anemia, unspecified
Disorders of magnesium metabolism
Polyneuropathy, unspecified
Dehydration
Other specified disorders of white blood cells
Hypokalemia
Thrombocytopenia, unspecified
Other specified anemias
B
pancreatic_cancer
Agranulocytosis secondary to cancer chemotherapy
Dehydration
Sepsis, unspecified organism
Iron deficiency anemia, unspecified
Other iron deficiency anaemias
Other specified disorders of white blood cells
Disorders of magnesium metabolism
Drug-induced polyneuropathy
Polyneuropathy, unspecified
Other specified anemias
Hypokalemia
Thrombocytopenia, unspecified
Hypotension, unspecified
Hypo-osmolality and hyponatraemia
C
STS-TE
Neoplasm related pain
Secondary malignant neoplasm of unspecified site
Dehydration
Constipation
Agranulocytosis secondary to cancer chemotherapy
D
STS-TE
Secondary malignant neoplasm of lung
Agranulocytosis secondary to cancer chemotherapy
Hypokalemia
Secondary malignant neoplasm of unspecified site
Neoplasm related pain
Constipation
Dehydration
Other specified anemias
E
STS-AR
Neoplasm related pain
Other specified anemias
Acute renal failure
Disorders of magnesium metabolism
Neutropenia, unspecified
Secondary malignant neoplasm of lung

Figure 7. Possible disease trajectories after pancreatic cancer, STS-TE, and STS-AR diagnoses. A) Trajectories with 2 hops after a pancreatic cancer diagnosis. B) Trajectories with 3 hops after a pancreatic cancer diagnosis. C) Trajectories with 2 hops after an STS-TE diagnosis. D) Trajectories with 3 hops after an STS-TE diagnosis. E) Trajectories with 2 hops after an STS-AR diagnosis. No trajectories with 3 hops after an STS-AR diagnosis were identified.

Trajectory validation

We systematically validated the significant disease trajectories with local data from the UTHealth OMOP CDM dataset to demonstrate the accuracy of our approach. Table 3 presents the validation results from the UTHealth OMOP CDM dataset for diagnosis pairs and long (i.e., 2- and 3-hop) trajectories identified in the IQVIA dataset. The results showed that 89.9%, 98.5%, and 83.9% of 1-hop trajectories found in the IQVIA cohorts were also found in the pancreatic cancer, STS-TE, and STS-AR patient cohorts of the UTHealth OMOP CDM, respectively. More long disease trajectories were found preceding the cancer diagnoses than following the cancer diagnoses in the UTHealth OMOP CDM cohorts. Specifically, trajectories with 2 hops preceding a pancreatic cancer, STS-TE, and STS-AR diagnosis were identified in 90.9% (vs. 46.3%), 83.3% (vs. 36.4%), and 41.2% (vs. 20%), respectively, of the UTHealth OMOP CDM cohorts. Longer disease trajectories (i.e., those with 3 hops) preceding and following cancer diagnoses were less covered. Trajectories with 3 hops preceding a pancreatic cancer or STS-TE diagnosis were covered by 22.8% and 8.1%, and trajectories with 3 hops following a pancreatic cancer or STS-TE diagnosis were covered by 10.1% and 5.9%, respectively, of patients in the UTHealth OMOP CDM dataset cohorts.

Table 3. Validation results of pairs with directionalities and long trajectories.

Cancer type	1-hop trajectory*	2-hop trajectory*	3-hop trajectory*

		Preceding cancer	Following cancer	Preceding cancer	Following cancer
Pancreatic	239/266 (89.9)	40/44 (90.9)	25/54 (46.3)	31/136 (22.8)	13/129 (10.1)
STS-TE	128/130 (98.5)	30/36 (83.3)	4/11 (36.4)	3/37 (8.1)	1/17 (5.9)
STS-AR	99/118 (83.9)	7/17 (41.2)	1/5 (20)	0/5 (0)	None

STS-AR, soft-tissue sarcoma of the abdomen and retroperitoneum; STS-TE, soft-tissue sarcoma of the trunk and extremity; UTHealth OMOP CDM, University of Texas Health Science Center Observational Medical Outcomes Partnership Common Data Model.

*All values are $n_{\text{UTHealth OMOP CDM}}/n_{\text{IQVIA}}$ (%).

DISCUSSION

In this study, we extracted signature disease trajectories from the IQVIA Oncology EMR for 3 rare cancers, which we subsequently validated using the UTHealth OMOP CDM dataset. Our identification of signature disease trajectories demonstrated the value of rendering real-world big data to better understand rare cancer disease patterns; 13,606 pancreatic cancer, 4,536 STS-TE, and 1,003 STS-AR patients in the IQVIA Oncology EMR and 2,568, 860, and 179 patients, respectively, in the UTHealth OMOP CDM were included in our study. This aggregation of rare cancer cases, along with the remaining 1.2 million cancer patients in the IQVIA oncology dataset that were included in the comparison groups, provided a rich and robust foundation for identifying signature disease trajectories. Our findings included 266 significant diagnosis pairs for pancreatic cancer, 130 significant pairs for STS-TE, and 118 significant pairs for STS-AR. In addition, we discovered 44 2-hop (i.e., 2-progression) and 136 3-hop trajectories before a pancreatic cancer diagnosis, 36 2-hop and 37 3-hop trajectories before an STS-TE diagnosis, and 17 2-hop and 5 3-hop trajectories before an STS-AR diagnosis. Meanwhile, we found 54 2-hop and 129 3-hop trajectories following a pancreatic cancer diagnosis, 11 2-hop and 17 3-hop trajectories following an STS-TE diagnosis, and 5 2-hop and 0 3-hop trajectories following an STS-AR diagnosis.

Our study offers 3 key contributions to understanding disease progression patterns. First, our approach is tailored to extract signature trajectories for a specific disease of interest. In our study, we identified distinct disease trajectories for 3 rare cancers. Notably, the two STS types exhibited some unique progression patterns. For example, essential (primary) hypertension (I10) progressed to other specified soft tissue disorders (M79.8) before a diagnosis of STS-TE (Table S7 of Supplementary Data 2, while this pattern was not significant in STS-AR (Table S12 of Supplementary Data 2). Mixed hyperlipidemia (E78.2) progressed to acute renal failure (N17.9), and dehydration (E86.0) before a diagnosis of STS-AR (Table S14 of Supplementary Data 2), while usually pain followed mixed hyperlipidemia before a diagnosis of STS-TE (Table S9 of Supplementary Data 2). Although we used the IQVIA Oncology EMR as the main data source, such an approach is generalizable to other diseases and real-world longitudinal patient records. Second, we systematically validated the discovered trajectories with local data, supporting the validity of the discovered trajectories. To our knowledge, this has not been done in previous studies. Third, the disease trajectories of pancreatic cancer, STS-TE, and STS-AR discovered in this study are potential resources for providers to deepen their understanding of the temporal progression of conditions preceding and following these rare cancers, further informing patient-care decisions.

Understanding disease trajectories is particularly critical in the rare cancer population, as identifying high-risk patients who may benefit from early surveillance or screening is challenging. Presently, high-risk surveillance for pancreatic cancer is typically reserved for patients with cystic lesions in their pancreas or those who have high familial and/or genetic risk.[16] Similarly, STS screening is reserved for patients with high-risk genetic alterations such as *TP53*, *RB1*, and *NF1* mutations. Our findings suggest that key clinical features and diagnoses such as hyperlipidemia, pain in the joints, urinary tract infection, disorders of magnesium metabolism, acute renal failure, and dehydration may precede the diagnoses of pancreatic cancer and STS that can potentially serve as early markers of disease.

Furthermore, some of the significant diagnosis pairs identified in our study have been reported in other studies. However, longer trajectories were not reviewed in the existing literature. For example, new-onset diabetes has been associated with a higher risk of pancreatic cancer.[19] A modest causal association between type 2 diabetes and pancreatic cancer was also established in a meta-analysis of 36 studies.[20] Another study using the UK Biobank data established a possible relationship between gastroesophageal reflux disease and pancreatic cancer.[21] Anemia, another diagnosis closely associated with pancreatic cancer in our study, has been found in 93% of patients presenting with pancreatic cancer.[22] Sarcoma could be misdiagnosed as chronic hypertension [23] and in some sarcoma patients, hyperlipidemia was found in their medical history.[24, 25]

The early signs and symptoms of STS are often challenging to differentiate from those of benign conditions. Patients with palpable masses associated with pain or growth are often recommended to undergo further evaluation.[26] However, patients with STS often experience significant delays in referral and treatment; one recent study suggested that patients can experience up to 14-month delays before a referral to a high-volume center.[27] In this same study, pain alone was not enough for a patient to meet the criteria for referral to a sarcoma center, despite earlier research listing it as one of 5 criteria for urgent screening.[28] Our data support these findings in earlier research, suggesting that pain in the joint or extremity may be an early marker of an STS-TE diagnosis. Similarly, STS-AR has vague signs and symptoms. Our 1- and 2-hop trajectories suggest that in patients with increasing abdominal girth or distention in conjunction with anemia and acute renal failure, clinicians should have a high index of suspicion for STS-AR.

Our work has the following limitations. First, the IQVIA dataset only spanned 8 years (2015 to 2023). If a patient was diagnosed with one of the diseases of interest after 2022, enough post-diagnosis data may not have been available from the 3-year period we used to calculate diagnosis pairs. Applying our method to longitudinal patient records over a longer duration is warranted to discover more complete disease trajectories. Second, we relied on one real-world dataset (UTHealth OMOP CDM) to validate the significant disease trajectories. Because this dataset is a single-center data source, data fragmentation may be present. Some patients may have had only intermittent records in the UTHealth OMOP CDM dataset, so diagnoses from admissions at other hospitals may not have been captured. However, even with this limitation, we successfully verified that most significant diagnosis pairs with directionalities for pancreatic cancer (89.9%), STS-TE (98.5%), and STS-AR (83.9%) existed in the respective UTHealth cohorts. Further validation of our method with multi-institutional datasets or other population-based datasets is warranted. Finally, the disease trajectories were made using ICD-10 codes that may not account for more granular clinical signs and symptoms, and our trajectories may not be completely accurate.

Conclusion

In this study, we proposed a method for discovering signature disease trajectories by leveraging large-scale EHR data. Validation with a local cohort confirmed the feasibility and reliability of the proposed method. Although we demonstrated the validity of our approach with only rare cancers and one real-world dataset, this approach is generalizable to other diseases and other real-world patient datasets. We plan to integrate the knowledge gained from this study to develop EHR-based predictive models to inform early diagnosis, treatment, and prognosis of pancreatic cancer, STS-TE and STS-AR.

Acknowledgment

Source: IQVIA Oncology EMR

We thank Madison Semro, Associate Scientific Editor, and Dawn Chalaire, Associate Director of Editing Services, in the Research Medical Library at The University of Texas MD Anderson Cancer Center for editing this article.

Funding: This project was supported by the Cancer Prevention Research Institute of Texas (CPRIT) RR230020, National Institute of Aging grant RF1AG072799, National Human Genome Research Institute R01HG12748, and National Library of Medicine R01LM11934.

Author Contributions:

L.W.: Conceptualized and designed the study, implemented algorithms, analyzed the data, and drafted the manuscript.

R.L.: Designed the study, implemented algorithms, analyzed the data, and drafted the manuscript.

A.W.: Analyzed the data and revised the manuscript.

Q.L.: Revised the manuscript.

J.W.: Revised the manuscript.

X.R.: Revised the manuscript.

C.L.R.: Revised the manuscript.

A.G.: Revised the manuscript.

N.M.: Revised the manuscript.

M.H.G.K.: Revised the manuscript.

Heather Lyu: Designed and supervised the study, analyzed the data, and revised the manuscript.

Hongfang Liu: Supervised the study and revised the manuscript.

Data Availability: The data that support the findings of this study are available from IQVIA but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of IQVIA. Other data is provided within the manuscript or supplementary information files.

References

1. NCI. Dictionary of Cancer Terms: “rare cancer” 2021 [cited 2024 September 4]. Available from: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/rare-cancer.

2. Reinke DK. Meaningful engagement of the patient in rare cancer research: sarcoma as an exemplar. Current Problems in Cancer. 2021;45(4):100772.

3. Stoffel EM, Brand RE, Goggins M. Pancreatic cancer: changing epidemiology and new approaches to risk assessment, early detection, and prevention. Gastroenterology. 2023;164(5):752-65.

4. Ilic I, Ilic M. International patterns in incidence and mortality trends of pancreatic cancer in the last three decades: A joinpoint regression analysis. World Journal of Gastroenterology. 2022;28(32):4698.

5. Zambo I, Veselý K. WHO classification of tumours of soft tissue and bone 2013: the main changes compared to the 3rd edition. Ceskoslovenska patologie. 2014;50(2):64-70.

6. Murray SA, Kendall M, Boyd K, Sheikh A. Illness trajectories and palliative care. Bmj. 2005;330(7498):1007-11.

7. Murtagh FE, Sheerin NS, Addington-Hall J, Higginson IJ. Trajectories of illness in stage 5 chronic kidney disease: a longitudinal study of patient symptoms and concerns in the last year of life. Clinical Journal of the American Society of Nephrology. 2011;6(7):1580-90.

8. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, Jensen PB, Jensen LJ, Brunak S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature communications. 2014;5(1):4022.

9. Hu JX, Helleberg M, Jensen AB, Brunak S, Lundgren J. A large-cohort, longitudinal study determines precancer disease routes across different cancer types. Cancer research. 2019;79(4):864-72.

10. Paik H, Kan MJ, Rappoport N, Hadley D, Sirota M, Chen B, Manber U, Cho SB, Butte AJ. Tracing diagnosis trajectories over millions of patients reveal an unexpected risk in schizophrenia. Scientific data. 2019;6(1):201.

11. Paik H, Kim J. Condensed trajectory of the temporal correlation of diseases and mortality extracted from over 300,000 patients in hospitals. Plos one. 2021;16(10):e0257894.

12. Siggaard T, Reguant R, Jørgensen IF, Haue AD, Lademann M, Aguayo-Orozco A, Hjaltelin JX, Jensen AB, Banasik K, Brunak S. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. Nature communications. 2020;11(1):4952.

13. Herzeel C, D’Hondt E, Vandeweerd V, Botermans W, Akand M, Van der Aa F, Wuyts R, Verachtert W. A software package for efficient patient trajectory analysis applied to analyzing bladder cancer development. PLOS Digital Health. 2023;2(11):e0000384.

14. Boman H, Hedelin H, Jacobsson S, Holmäng S. Newly diagnosed bladder cancer: the relationship of initial symptoms, degree of microhematuria and tumor marker status. The Journal of urology. 2002;168(5):1955-9.

15. DeGEORGE KC, Holt HR, Hodges SC. Bladder cancer: diagnosis and treatment. American family physician. 2017;96(8):507-14.

16. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, Dymshyts D, Hripcsak G. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for

international data harmonization. Journal of the American Medical Informatics Association. 2024;31(3):583-90.

17. Weisstein EW. Bonferroni correction. https://mathworld wolfram com/. 2004.

18. Bastian M, Heymann S, Jacomy M, editors. Gephi: an open source software for exploring and manipulating networks. Proceedings of the international AAAI conference on web and social media; 2009.

19. Yuan C, Babic A, Khalaf N, Nowak JA, Brais LK, Rubinson DA, Ng K, Aguirre AJ, Pandharipande PV, Fuchs CS. Diabetes, weight change, and pancreatic cancer risk. JAMA oncology. 2020;6(10):e202948-e.

20. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, Barzi F, Woodward M. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. British journal of cancer. 2005;92(11):2076-83.

21. Wu G, Liu Y, Ning D, Zhao M, Li X, Chang L, Hu Q, Li Y, Cheng L, Huang Y. Unraveling the causality between gastroesophageal reflux disease and increased cancer risk: evidence from the UK Biobank and GWAS consortia. BMC medicine. 2024;22(1):323.

22. Tchekmedyian NS. Anemia in cancer patients: significance, epidemiology, and current therapy. Oncology (Williston Park, NY). 2002;16(9 Suppl 10):17-24.

23. Widera E, Sulica R. Pulmonary artery sarcoma misdiagnosed as chronic thromboembolic pulmonary hypertension. Mount Sinai Journal of Medicine. 2005;72(6).

24. Ahmed A, Amin S, Aljumaili Z, Elsayad M, Thomas-Ogunniyi J. A Rare Case with KMT2D: BCOR Fusion Alteration Sarcoma in A 14-Year-Old-Female. American Journal of Clinical Pathology. 2024;162(Supplement_1):S9-S10.

25. Patel H, Gregg J, Matthews D. Pleural Spindle Cell Sarcoma Masquerading as a Rapidly Growing Pulmonary Infection. B80-7 DIAGNOSTIC DILEMMAS IN THORACIC ONCOLOGY: American Thoracic Society; 2024. p. A4635-A.

26. Dangoor A, Seddon B, Gerrand C, Grimer R, Whelan J, Judson I. UK guidelines for the management of soft tissue sarcomas. Clinical sarcoma research. 2016;6:1-26.

27. Pencavel TD, Strauss DC, Thomas GP, Thomas JM, Hayes AJ. Does the two-week rule pathway improve the diagnosis of soft tissue sarcoma? A retrospective review of referral patterns and outcomes over five years in a regional sarcoma centre. The Annals of The Royal College of Surgeons of England. 2010;92(5):417-21.

28. George A, Grimer R. Early symptoms of bone and soft tissue sarcomas: could they be diagnosed earlier? The Annals of The Royal College of Surgeons of England. 2012;94(4):261-6.