Ma *et al. Archives of Public Health* (2025) 83:100
https://doi.org/10.1186/s13690-025-01589-1

Archives of Public Health

RESEARCH Open Access

Discovering sequential patterns and interrelations among multiple diseases in electronic medical records using cSPADE algorithm

He Ma[1,2], Qianxin Huang[3], Hong Zhang[4], Hui Song[5], Bo Zhang[6], Ying Liu[2] and Lin Zhang[1*]

Abstract

Background The intricate relationships between diseases are characterized by the sequence and temporal intervals of their onset, which are critical for understanding the essence of comorbidity and predicting disease progression. This study seeks to investigate the interdependencies and chronological order of various diseases that occur in the same patient by employing sequential pattern mining algorithms. Specifically, the research endeavors to delineate the disparities in the time intervals between the onset of distinct disorders and to scrutinize the concordance and discordance in disease sequence patterns across gender groups.

Methods Patient identity information, visit dates, and diagnostic data were aggregated from the electronic medical record databases of three large general hospitals. The diagnostic information included the International Classification of Diseases, Tenth Revision (ICD-10) codes, along with their corresponding descriptions. A total of 1,060,344 diagnostic entries from 269,973 patients who visited during 2012–2022 were incorporated into the mining model, which was constructed using the Sequential Pattern Discovery using Equivalence Classes (SPADE) algorithm.

Results A total of 212 highly supported sequential pattern rules were ultimately identified, most of which were related to disorders of the endocrine and circulatory systems. In 66 patterns, the order of disease incidence or diagnosis was relatively well-defined. The time interval between the onset of two diseases ranged from 1 to 2 years in most patterns. For patterns with short-term relationships, the interval was less than 2 months, whereas in some cases, the interval extended to 5 to 10 years. Among the extracted patterns, 176 exhibited stronger support in the male dataset compared to the female dataset. Patterns related to cardiovascular and liver diseases were more prevalent in males, while those associated with orthopedic and endocrine disorders showed higher prevalence in females.

Conclusion Our findings demonstrate the effectiveness of the constrained SPADE (cSPADE) algorithm in comorbidity research and highlight several clinically significant sequential comorbidity patterns. These patterns are expected to contribute to disease prevention, etiological research, and the development of clinical decision support systems.

Keywords cSPADE, Sequential pattern mining, Comorbidity

*Correspondence:
Lin Zhang
lin.zhang@cumt.edu.cn

Full list of author information is available at the end of the article

Text box 1. Contributions to the literature

- This study employs the SPADE algorithm to uncover temporal sequences between comorbid diseases, offering novel insights into disease progression.
- It highlights significant gender differences in comorbidity patterns, enabling more tailored prevention and intervention efforts. These findings help healthcare providers design specific programs and treatments that cater to the unique health needs of different genders.
- The research enhances the understanding of different time intervals between the onsets of diseases, facilitating more accurate disease prediction and optimized management strategies. This knowledge allows for better anticipation of patient needs and the timely implementation of early interventions.

Introduction

The intricate and multifaceted relationships among various disorders are of great significance. This complexity often manifests as the co-occurrence of multiple diseases or medical conditions within a single patient, exceeding the likelihood of random chance. It is common for two or more conditions to be underpinned by shared risk factors or for one disease to precipitate another. Investigating the patterns of inter-disease interaction is essential. Gaining insights into the pathogenesis of diseases necessitates a thorough understanding of how these conditions may influence and intersect with one another. This entails a detailed analysis of the biological networks and pathways potentially disrupted by successive or concurrent pathologies, thus improving our understanding of the etiological contributions to disease expression. Thus, identifying comorbidity patterns equips healthcare providers with the foresight to predict potential complications and develop personalized treatment strategies. It is especially crucial when multiple conditions may interact synergistically to worsen patient outcomes, necessitating a holistic and integrated therapeutic approach. Moreover, a better understanding of the interrelationships among diseases can enhance diagnostic precision. The exploration of disease interconnections contributes to the knowledge base of Clinical Decision Support Systems (CDSS). Identifying shared pathways among diseases will enable healthcare systems to allocate resources more effectively, streamline patient care pathways, and reduce the overall strain on healthcare infrastructure. This approach may help bolster population health and ensure equitable access to healthcare services.

Hence, exploring patterns of comorbidity has consistently been a key focus of academic research. Several scholars [1–9] have explored the interrelationships among disorders at multiple levels, including population, cellular, and molecular perspectives, providing significant insights. However, current research generally draws on clinical observations and experiences, with a predominant focus on common and frequently occurring diseases. Moreover, existing studies were often confined to specific types of disorders or particular bodily systems, lacking a comprehensive, macro-level examination of comorbidity relationships. This oversight may fail to account for the interactions between seemingly "unrelated" diseases. Thus, Jensen [10] conducted a relatively comprehensive analysis in Denmark to identify temporal disease trajectories. However, it may not be generalizable to other populations, such as Asian populations, with different healthcare systems, lifestyles, or genetic predispositions. Furthermore, since the study was conducted 10 years ago, it did not consider the recent advancements in diagnostics or treatment that might alter the observed disease trajectories.

In recent years, the widespread implementation of hospital information systems, electronic medical records (EMR), and government-built big data health platforms has led to the accumulation of health-related data. This surge in data availability has catalyzed the prominence of big data analytics, such as machine learning, in medical research. Tasks once deemed unattainable through traditional methods have become feasible. For instance, during efforts to curb the widespread transmission of Corona Virus Disease 2019 (COVID-19), advanced big data techniques, including convolutional neural networks and metaheuristic algorithms, played a pivotal role in refining public quarantine policies, optimizing the allocation of medical resources, and forecasting disease trends [11]. Innovative digital contact tracing technology has also accurately predicted the trajectories of viral spread while safeguarding individual privacy, contributing significantly to the containment of the virus [12]. Similarly, in comorbidity research, techniques such as association rule mining and sequential pattern mining are employed to extract insights from large datasets, offering the potential to uncover novel clues and deepen the understanding of complex disease relationships.

We employed the Apriori algorithm to mine an EMR dataset and identified 110 association rules [13]. However, association rule mining could only demonstrate the existence of correlations between two diseases and failed to elucidate the temporal sequence of disease occurrence, a critical requirement for inferring causality. Additionally, our previous study was conducted in a single-center environment, which constrained the representativeness of the evidence gathered. Limited by the available resources and circumstances at the time, the influence of gender on inter-disease relationships was not assessed. Therefore, our knowledge remains markedly limited regarding the causal and directional relationships, particularly among Asian populations.

Directional relationships refer to situations in which some diseases are sequentially associated, with the onset of a primary disease preceding that of other conditions

or stages. In this context, sequence pattern mining algorithms gain substantial relevance. Sequence pattern mining algorithms are a class of data mining techniques that aim to identify frequent and meaningful subsequences within a dataset of ordered events or transactions. These algorithms are instrumental in bioinformatics [14], performance analysis in competitive sports [15], animal behavior analysis [16], and opinion mining [17], where the order of events is crucial for understanding patterns and making predictions. Analogous to the Apriori algorithm, sequence pattern mining algorithms are adept at identifying interrelationships among diseases and quantifying the strength of their associations. However, a key distinction is that sequence pattern mining algorithms elucidate the order and timing of disease occurrences, providing a more nuanced understanding of disease progression pathways and bolstering the evidence required for etiological inference. Furthermore, the patterns unearthed by sequence pattern mining algorithms are more practical and beneficial in CDSS construction and disease prediction.

Over the years, sequential pattern mining has diversified into various specific algorithms, each with unique merits. Among these, the Sequential Pattern Discovery using Equivalence Classes (SPADE) algorithm [18] stands out due to its lattice-theoretic approach, which constructs equivalence classes to share common prefixes among sequences. This innovation significantly enhances support counting efficiency by minimizing the need for repeated database scans, a frequent performance bottleneck in other algorithms. Additionally, SPADE inherently supports parallelism, leveraging multicore and multiprocessor environments by computing frequencies independently for each equivalence class. This robust design facilitates the efficient extraction of sequential patterns, particularly in large databases, by optimizing computational resources and supporting parallel execution. SPADE is highly versatile, capable of handling complex temporal constraints and adapting to diverse sequence mining tasks, such as mining therapeutic pathways for breast cancer [19], development characteristics of pediatric obesity [20], and cancer co-occurrence patterns [21].

This investigation was a retrospective study designed to leverage the SPADE algorithm to extract temporal information and sequential patterns pertaining to distinct diseases occurring successively in the same patient. These patterns can contribute to a comprehensive understanding of the macro network of directional inter-disease relationships among Asian populations and may uncover associations between certain disorders that were previously considered "unrelated". Furthermore, sequence patterns from different gender subgroups were investigated to evaluate the impact of gender on inter-disorder relationships. We aimed for the sequential patterns mined not only to offer novel insights for etiological researchers but also to be integrated into disease prediction models and CDSS.

Materials and methods

The schematic diagram of this study is provided in Fig. 1.

Data source

In contrast to our previous research [13], in this study, we employed data from three comprehensive tertiary-level hospitals at grade A: the Affiliated Hospital of Xuzhou Medical University, Xuzhou First People's Hospital, and Northern Jiangsu People's Hospital. These hospitals are located in two major cities, Xuzhou and Yangzhou, in Jiangsu Province, China, which bolsters the robustness of the evidence. Annually, millions of outpatients and inpatients seek medical treatment at these three hospitals, predominantly coming from more than 160 cities throughout China, thereby enhancing the representativeness and persuasiveness of the study's findings.

The raw data for this study were derived from the outpatient and inpatient medical records in the EMR databases of the aforementioned three hospitals. The prerequisite for conducting sequential pattern mining was to discern multiple hospital admissions of the same patient and establish their chronological order of visits and corresponding medications. Thus, we extracted basic information such as patients' names, gender, dates of birth, and national identification numbers to distinguish individual patients. The national identification number, which serves as the official identifier for each citizen of China, was prioritized when it was fully available for identifying patient identity. In instances where the national identification number was missing, such as in emergency admissions or due to deficiencies in EMR quality control, a combination of name, gender, and date of birth was used to distinguish participants. In addition to the identity information mentioned above, the date of each outpatient visit was used to determine the chronological order of visits by the same patient. Conversely, when a patient was admitted through inpatient services, the date of admission was utilized. Furthermore, date data were employed to calculate the time intervals between multiple medical consultations, reflecting the span between the occurrences of different diseases within the same sequence. All information was anonymized to ensure patient privacy by re-encoding participants' national identification numbers or the combination of names and dates of birth. Specifically, each patient was assigned a unique identifier to replace the original identity information for inclusion in the analysis model. Since the study did not involve human experimentation, it was exempted from ethical approval by the Ethics Committee of the Affiliated Hospital of Xuzhou Medical University.

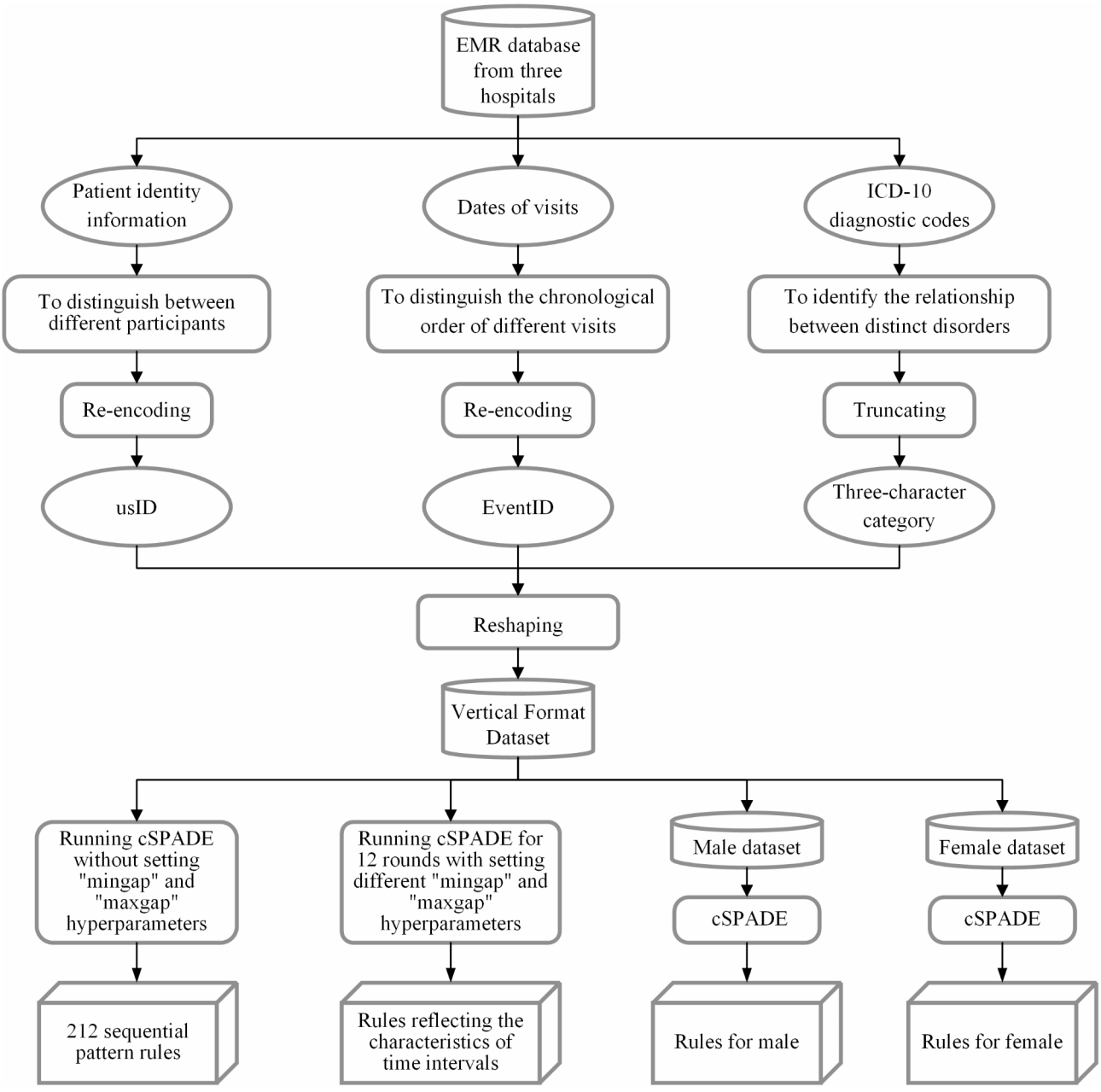

Fig. 1 Flowchart illustrating the data mining process of the interrelations among multiple diseases

The focal point of sequential pattern mining pertains to the diagnostic data from each patient encounter. These diagnostic data were predominantly represented in the database and model through codes from the International Classification of Diseases, Tenth Revision (ICD-10), along with textual descriptions. ICD-10, developed by the World Health Organization (WHO), is a coding system designed to facilitate the systematic recording, analysis, and comparison of global mortality and morbidity data. ICD-10 serves as a standard diagnostic tool for epidemiology, health management, and clinical purposes. The structure of the ICD-10 coding system is organized into chapters, each addressing a different type of disease or health condition. ICD-10 codes can be divided into "categories" and "subcategories" depending on the level of detail in the disease descriptions. Although subcategories provide a more detailed delineation of disease entities, the three-character category codes generally hold greater statistical relevance. They are used more often in international reporting and comparative analysis. Therefore, the three-character category was incorporated into the mining model for analysis. Furthermore, "Z" codes in ICD-10 represent situations that do not arise from an illness or injury but influence current health status or

necessitate healthcare services. Additionally, "V", "W", "X", and "Y" codes capture a variety of external causes of morbidity and mortality. Hence, "Z", "V", "W", "X", and "Y" codes were excluded, having no significant impact on the operational efficacy of the model, nor did they alter the analytical outcomes.

Professional coders constructed ICD-10 codes by extracting key information from medical documents written by clinicians, and these codes were then entered into the EMR database. All coders involved in this process had at least three years of professional experience. To enhance accuracy and efficiency, the EMR system was equipped with integrated search tools that assisted coders in identifying the most appropriate codes. Each medical record was independently coded by two coders to ensure consistency and comparability. In cases where discrepancies arose between the two sets of codes, an experienced senior coding supervisor reviewed the differences, consulted with the patient's attending physician, and made the final determination. Through this rigorous workflow and quality control process, the integrity and precision of coding were reliably maintained.

Some records were excluded from this investigation: (1) patients with only a single hospitalization during the study period; (2) diagnostic information containing ICD-10 "V", "W", "X", "Y", "Z" codes or codes with corresponding descriptions including the word "other"; (3) diagnostic information containing ICD-10 "R" codes since they primarily represent a wide range of symptoms, signs, and observations insufficiently specific to indicate a particular disease or health condition; (4) when encountering records with missing values or logical errors, an attempt was made to infer and fill in missing information or rectify inaccuracies through analysis of redundant data. If rectification proved unfeasible, such records were excluded.

Moreover, medical records from January 1, 2012, to December 31, 2022, were included based on data quality and availability. Finally, 1,060,344 clinical records related to 269,973 patients were integrated into the model, encompassing 2,491,690 primary and secondary diagnostic codes.

Table 1 Sample of the structured dataset used in the cSPADE data mining model

usID	eventID	Diagnostic code 1	Diagnostic code 2	Diagnostic code 3	…
1	1998	J18	J93	D64	
1	2024	J18	J93	D75	
2	501	F48			
2	1802	M51	M48		
3	1534	G35			
3	1745	G35			
3	1898	I70			
3	2165	G36			
4	1231	J21			
4	2001	J18			
4	2070	J03	B96		
5	2588	G40	N20		
5	2614	G40			
6	3462	C83	I31		
6	3508	C83	I31	D61	
7	2515	K35			
7	2521	K35			

Data preparation

SPADE outperforms other sequence pattern mining algorithms, such as GSP and PrefixSpan, in terms of execution time, particularly when precomputing the support of 2-sequences. SPADE's performance is an order of magnitude faster than GSP, while also demonstrating excellent scalability when applied to large-scale datasets. Although PrefixSpan outperforms SPADE in certain scenarios, SPADE exhibits superior performance when applied to shorter sequences [22, 23]. Given the large dataset in this study and the short sequential relationships observed between diseases, the SPADE [18] algorithm was employed for sequence pattern mining.

First, we generated a unique sequence identifier (usID) for each patient based on his or her national identification number or a combination of their name, gender, and date of birth. This procedure anonymized identity information and allowed the algorithm to distinguish between individual patients. Additionally, the number of days between each admission or visit date and January 1, 2012, was calculated to generate the eventID, which was used to distinguish medical encounters for the same individual (for example, "eventID = 1998" signifies that the patient's consultation occurred 1998 days after January 1, 2012, which corresponds to June 20, 2017.) and to compute the temporal interval between the onset of two diseases (for instance, if a patient is associated with two eventIDs, 1998 and 2024, it indicates a 26-day interval between the two consultation dates.). Subsequently, each original diagnostic code was truncated following the requisites of the ICD-10 category. Finally, all diagnostic codes from a single visit were combined into a single row, while codes from different visits were vertically sequenced in chronological order according to the date of consultation or admission. This arrangement resulted in the dataset used in the model. An illustrative example of the dataset is shown in Table 1.

Data analysis

The SPADE algorithm initially generated a candidate sequence dataset. In the candidate dataset, the consultation dates and diagnostic information of each patient formed a "sequence", sometimes referred to as "patterns". For example, the sequence "I10 → I63" indicated that the

patient presented with essential hypertension followed by cerebral infarction, indicating that essential hypertension preceded cerebral infarction.

Once the candidate dataset was organized, SPADE identified frequent items and sequences by calculating their support, which is defined as the number of data sequences in which a particular sequence appears. In the SPADE algorithm, the term "support" refers to a measure quantifying the prevalence of patterns within a given dataset. The support value is typically expressed as a percentage or as a fraction of the total number of sequences, and it is used to determine whether a sequence is sufficiently common to be considered significant or interesting. In the context of sequence mining, a pattern has high support if it occurs frequently across the dataset. Conversely, low support indicates that the pattern is relatively rare. For instance, if one examines the pattern "I10" (Essential hypertension) followed by "I63" (Cerebral infarction), the support for this sequence is 0.006086. This value is calculated by dividing the number of transactions containing this exact sequence (1643) by the total number of transactions (269973) in the dataset. The SPADE algorithm relies on the concept of support to prune the search space by eliminating sequences with support values below a user-defined threshold, often known as the minimum support threshold. Thus, setting a minimum support threshold is a crucial step. Unfortunately, the academic community lacks consensus on how to establish a standardized minimum support threshold. Excessive or insufficient thresholds each have advantages and drawbacks. Lowering the threshold may result in a larger number of frequent sequences, including potentially noisy or less significant patterns. Conversely, setting the threshold too high may exclude interesting patterns that do not occur as frequently. With a fixed "maxsize = 1 and maxlen = 2", we conducted multiple experimental iterations to evaluate the most suitable minimum support threshold. When the minimum support threshold was set to 0.01, the algorithm successfully identified 229 sequential patterns. Reducing the threshold to 0.005 led to the discovery of 527 patterns, whereas a further reduction to 0.001 resulted in the extraction of 2736 patterns. Finally, setting the support threshold as low as 0.0001 allowed the identification of 21,805 patterns. According to the World Health Organization, diseases with a prevalence of lower than 0.1% are considered rare [24]. We drew upon this definition as a reference to maximize the identification of disease relationships with relatively low incidence rates. Balancing this objective with the workload for manual validation by clinical experts within our research team, we ultimately set the minimum support threshold at 0.001. Subsequent to the execution of the algorithm, patterns exhibiting support less than 0.001 were automatically expurgated. On this basis, the ultimate patterns and conclusions were elicited after evaluation by clinical experts, ensuring alignment with fundamental medical axioms.

All frequent sequence patterns were mined using "***arulesSequences***", an R package designed for the constrained SPADE (cSPADE) algorithm. The cSPADE algorithm is an extension of the original SPADE algorithm, providing additional flexibility by allowing users to specify conditions that the discovered sequences must meet. It can significantly reduce the search space and focus the mining process on patterns that are relevant to specific research questions or business needs. In other words, cSPADE offers a specialized tool for finding patterns that are frequent and conform to pre-established constraints. This provides a more tailored approach to sequential pattern mining. Despite the minimum support threshold, these constraints in cSPADE could include, but are not limited to, the following: maxsize (an integer value specifying the maximum number of items of an element of a sequence), maxlen (an integer value determining the maximum length of the sequences to be considered. A sequence's length is the number of itemsets it contains. Limiting the length can reduce computation time by ignoring longer, potentially less significant patterns), maxgap (an integer value specifying the maximum time difference between consecutive elements of a sequence, defining the maximum "distance" between itemsets within a sequence for them to be considered contiguous), mingap (an integer value specifying the minimum time difference between consecutive elements of a sequence, forcing a minimum distance between itemsets to be maintained in the resulting patterns.). Given the interpretability of the analysis results and the pragmatic utility of the conclusions, the parameter "maxsize" was constrained to 1, whereas "maxlen" was set to 2 within the context of this investigation. We performed multiple iterations of the cSPADE algorithm with varied configurations of the "mingap" and "maxgap" parameters to conduct subgroup analyses. The durations between admission dates of two consecutive itemsets (diseases) in a sequence pattern were set as 12 temporal windows as follows: "≤ 1 week", "> 1 week and ≤ 2 weeks", "> 2 weeks and ≤ 1 month", "> 1 month and ≤ 2 months", "> 2 months and ≤ 3 months", "> 3 months and ≤ 6 months", "> 6 months and ≤ 1 year", "> 1 year and ≤ 2 years", "> 2 years and ≤ 3 years", "> 3 years and ≤ 5 years", "> 5 years and ≤ 7 years" and "> 7 years and ≤ 10 years", respectively, when executing the algorithm. This approach aimed to explore the precise temporal intervals between the onset of disparate diseases, maximizing the temporal characteristics of the algorithm. Additionally, we discussed the disparities in the "support" magnitudes of the same pattern between different genders to provide evidence for the enhanced precision of disease prediction.

The data were stored in the format of "Comma-Separated values" (.csv) and extracted from the EMR database constructed based on ORACLE 10 g (Integrated Development Environment: PL/SQL Developer 7.1.5.1399). Base R (ver.4.3.2) and the "***reshape2***" (ver.1.4.4) package were used for the data preprocessing. The "***arulesSequences***" (ver.0.2–30) package was employed to execute the cSPADE algorithm.

Results

Demographic information

Within the cohort of 269,973 subjects, the gender distribution included 132,575 (49.11%) female participants and 137,398 (50.89%) male participants. The age of patients at their first recorded clinical visit ranged from 1 day to 104 years, with a median age of 54 years. As shown in Fig. 2, aside from a notably high prevalence of diseases among children aged 0 to 10 years, the age distribution exhibited a slight skew. The largest proportion (111375 cases, 41.25%) of patients fell within the 50 to 70-year age range, making this demographic the most significant subset of the study population. All patients were treated at least twice in the three hospitals, with the maximum number of visits by a single patient during the study period reaching 154. The interval between the initial and final clinical visits was determined based on the eventID assigned to each visit. In this study, the minimum interval recorded between two visits was 1 day, while the maximum reached 4007 days (approximately 10.97 years). The median interval was 553 days (approximately 1.52 years). As depicted in Fig. 3, the frequency of visit intervals decreased progressively as the time span lengthened. 47.95% (129463 cases) of patients experienced an interval of less than 500 days between two visits. This observation suggested that short-term interactions among diseases were more prevalent than long-term relationships.

Among the 2,491,690 entries of the ICD-10 diagnostic code, a statistical analysis based on the initial alphabetic designation (corresponding to the "chapters" in ICD-10) revealed that "Diseases of the Circulatory System" ("I" codes) were the most prevalent, with 687,555 instances, accounting for 27.59% of the total. Following this are "Neoplasms" ("C" codes, 16.99%), and "Endocrine, Nutritional and Metabolic Diseases" ("E" codes, 9.17%) (Fig. 4). Further examination based on ICD-10 categories (first three characters) indicated that the three leading diseases in terms of prevalence were "Essential (primary) hypertension" (I10, accounting for 9.15%), "Cerebral infarction" (I63, constituting 5.7%), and "Type 2 diabetes mellitus" (E11, representing 4.97%), as shown in Fig. 5.

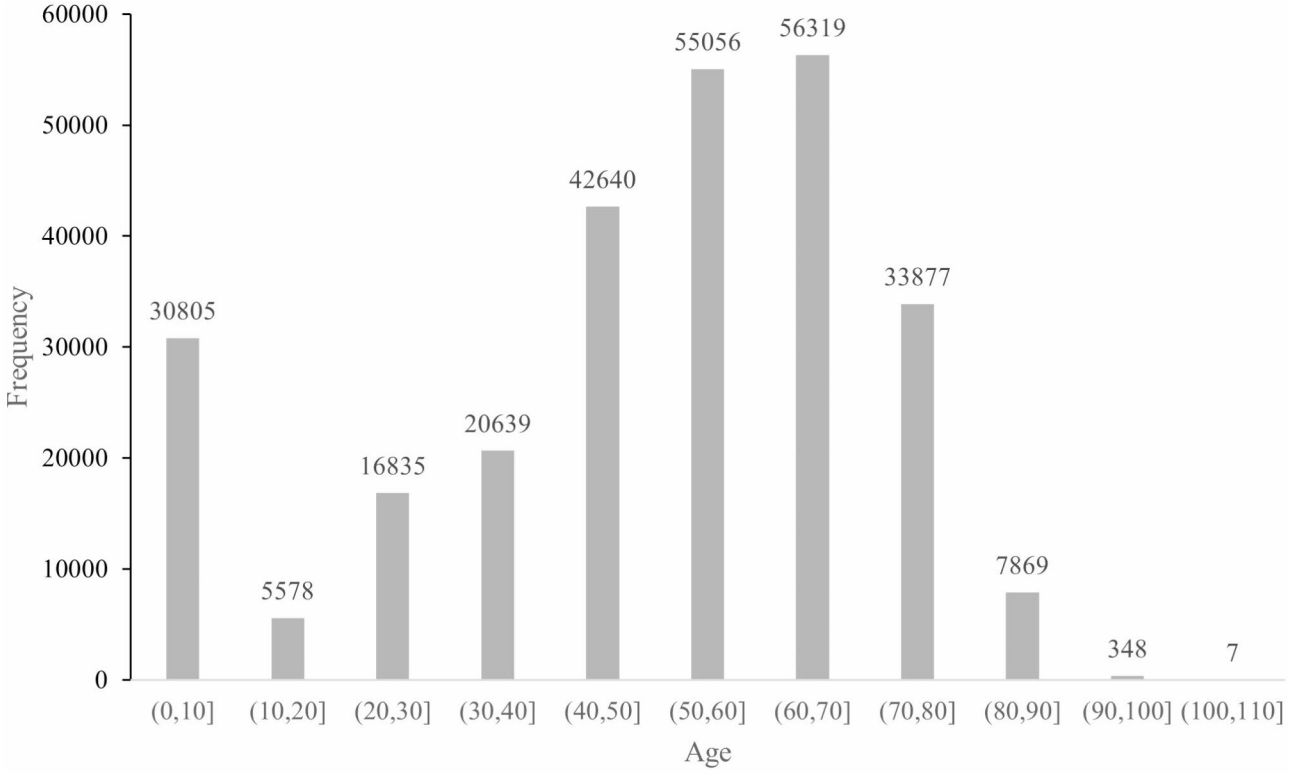

Fig. 2 Age distribution of patients included in the literature

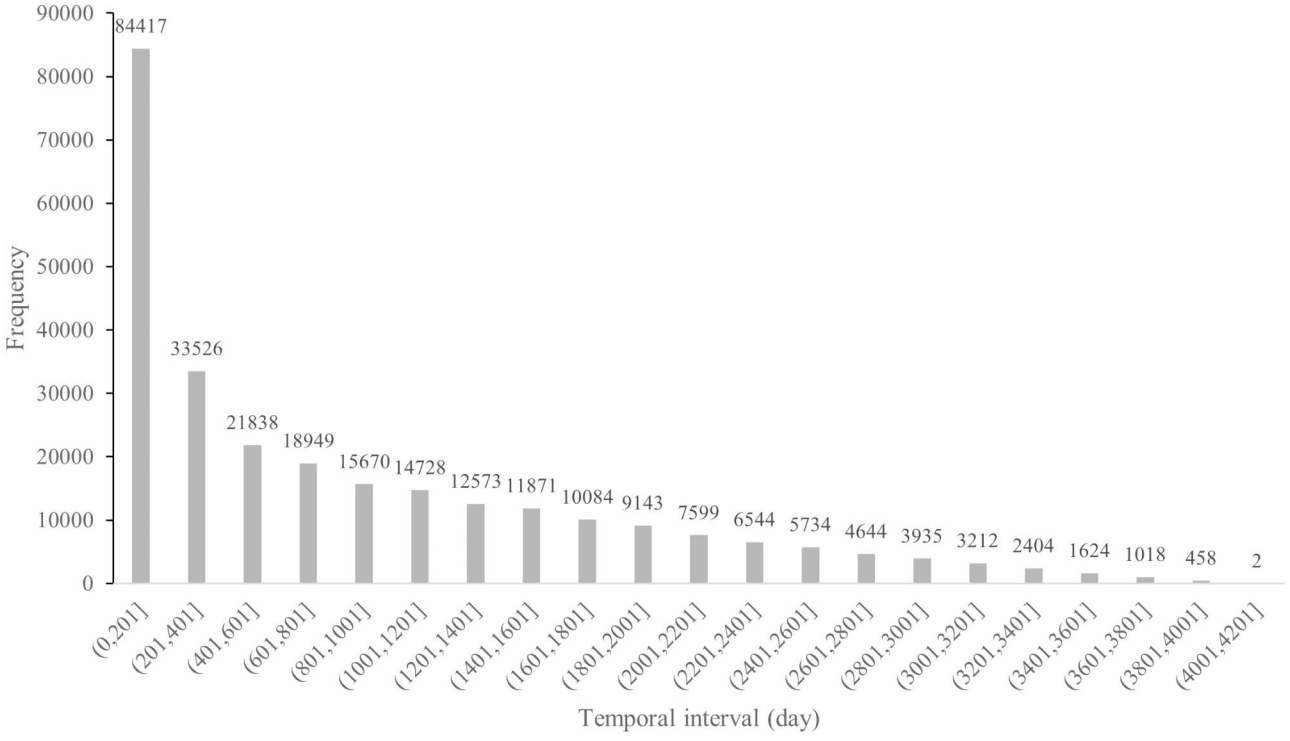

Fig. 3 Temporal interval distribution between first and last hospital consultations of patients in the study

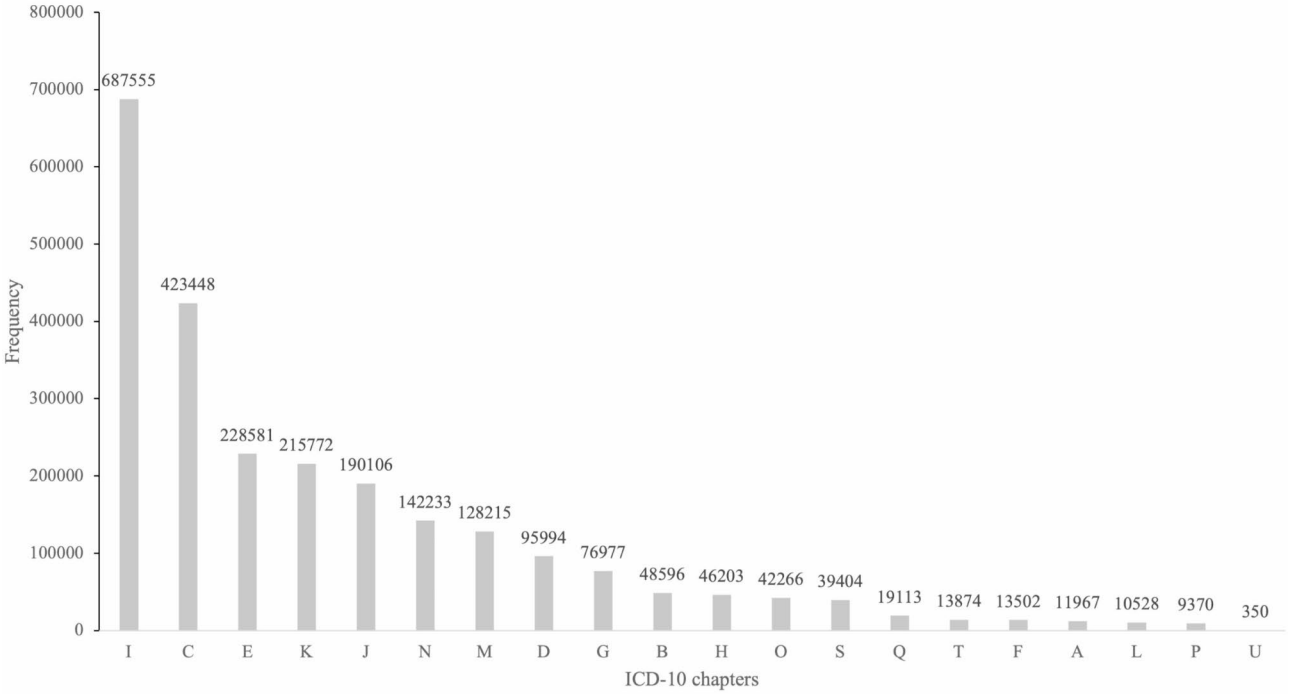

Fig. 4 Distribution of diagnostic codes for patients in the study (classified by ICD-10 "chapters")

The general state of the rules mined by the cSPADE algorithm

After configuring all the hyperparameters to run the cSPADE algorithm and manually screening each instance, 212 sequential pattern rules with support greater than 0.001 were ultimately included in the analysis. In other words, each rule was supported by at least 270 individuals. If diseases with relatively earlier medical visits were

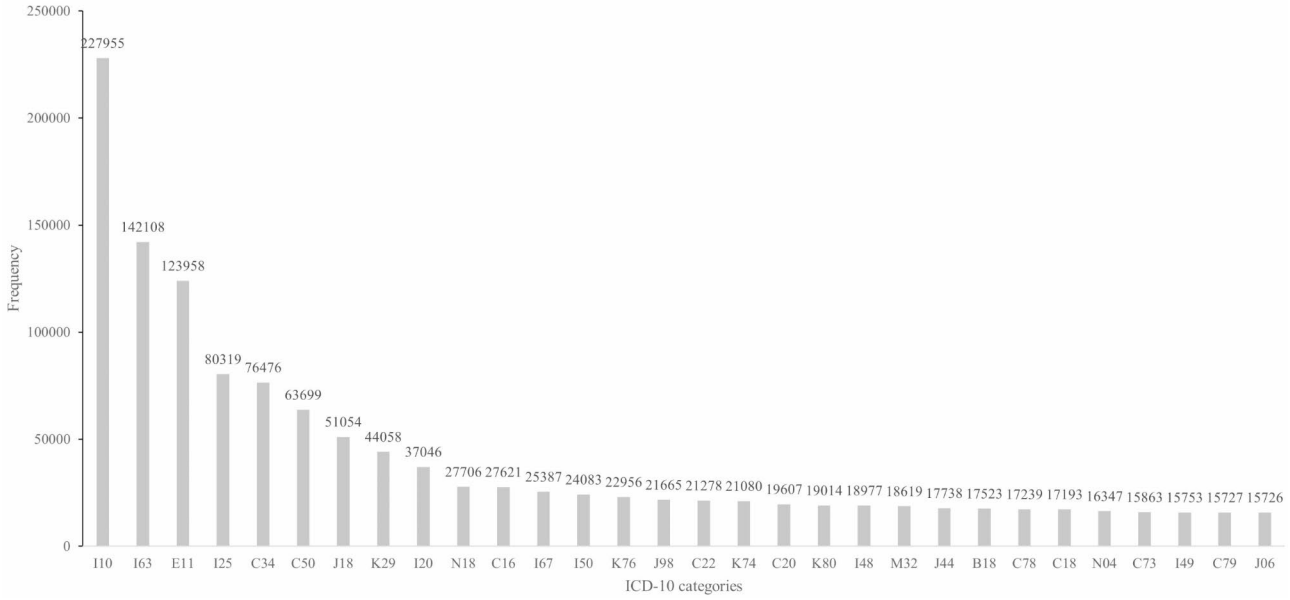

Fig. 5 Distribution of top 30 diagnostic codes for patients in the study (classified by ICD-10 "categories")

termed "antecedents" and those with later visits were referred to as "consequents", among these 212 rules, 146 maintained support above 0.001 even after interchanging the antecedents and consequents. This indicated within these 146 rules, the order of disease occurrence was not rigidly fixed. Certainly, the support for the newly formed rules after the interchange did not equate to that of the original rules, suggesting that one of the two diseases still tended to precede the other in occurrence or diagnosis. For instance, the rule "K29 → I63" indicated that patients with "Gastritis and duodenitis" might be prone to subsequent "Cerebral infarction", with support of 0.0038, indicating that within this study, 1028 patients experienced "Gastritis and duodenitis" prior to suffering a "Cerebral infarction". Conversely, the rule "I63 → K29" possessed a support level of 0.0026, indicating that 690 patients developed "Gastritis and duodenitis" after "Cerebral infarction". When synthesizing these two rules, the support levels suggested that "Gastritis and duodenitis" was more likely to occur before a "Cerebral infarction". In contrast to the aforementioned, the remaining 66 rules exhibited a more definitive sequence of disease occurrence or patient consultations. The corresponding rules did not exist or were removed by the algorithm because the support dropped below 0.001 after interchanging the antecedent and consequent diseases.

First, let us delve into the 66 rules that exhibit a comparatively definitive sequence in the occurrence or diagnosis of diseases (see Additional file 1). Forty-six of the rules indicated that the antecedent and consequent disease belonged to distinct body systems (as indicated by the first letter of the ICD-10 codes). This suggested the possibility of cross-systemic connections between disparate diseases. Continuing the analysis, 35 rules featured antecedent diseases that stemmed from the circulatory system (I codes), followed by eight rules in which the antecedent diseases pertained to endocrine and metabolic diseases (E codes). Twenty-nine rules had consequent diseases that also originated from the circulatory system (I codes), with the number of rules featuring diseases of the digestive system (K codes) as consequents ranking second. To be more specific, the rules with "Essential (primary) hypertension" (I10) and "Type 2 diabetes mellitus" (E11) as antecedents were the most frequent, whereas the number of rules with "Cholelithiasis" (K80) and "Heart failure" (I50) as consequents also exceeded those compared to other diseases. Within these 66 rules, "I25 → I20" possessed the highest support, indicating that 8196 individuals developed "Angina pectoris" after being diagnosed with "Chronic ischaemic heart disease". "I10 → I50" had the second position in support, revealing that 5379 patients, following their initial consultation for "Essential (primary) hypertension", subsequently sought medical care for "Heart Failure". These rules depicted a directed network structure concerning the relationships between diseases. For instance, "Essential (primary) hypertension" (I10) served as an antecedent for 14 categories of diseases, including "Heart failure" (I50), "Sequelae of cerebrovascular disease" (I69), and "Atrial fibrillation and flutter" (I48). Conversely, "Essential (primary) hypertension" was the consequent of three types of diseases: "Chronic nephritic syndrome" (N03), "Nephrotic syndrome" (N04), and "Chronic kidney disease" (N18). In addition, apart from "Essential (primary) hypertension" (I10), "Chronic ischaemic heart disease" (I25), "Type 2 diabetes mellitus" (E11), and "Angina

pectoris" (I20) could also act as antecedents for "Heart failure" (I50). This scenario illustrated the complexity of the network of relationships among diseases, thus increasing the complexity of analysis.

When our focus shifted to the remaining 146 rules where the positions of antecedent and consequent were interchangeable, 116 rules were identified in which the antecedents and consequents belonged to different bodily systems (see Additional file 2). "Diseases of the circulatory system" (I codes, 148 times) were the most frequently occurring conditions within the antecedents or consequents, followed by "Endocrine, nutritional and metabolic diseases" (E codes, 30 times) and "Malignant neoplasms" (C codes, 26 times). From the perspective of the ICD-10 category, "Essential (primary) hypertension" (I10), "Cerebral infarction" (I63), and "Type 2 diabetes mellitus" (E11) ranked among the top three diseases as antecedents or consequents. Within the 146 identified rules, "E11 → I10" exhibited the highest support of 0.067703067, representing 18,278 patients who, after being diagnosed with "Type 2 diabetes mellitus", sought medical attention for "Essential (primary) hypertension". However, only 450 patients were diagnosed with "Type 2 diabetes mellitus" (E11) after being diagnosed with "Essential (primary) hypertension" (I10) upon reversing the antecedent and consequent. This alteration rendered the support for the rule "I10 → E11" merely 0.001666833. These two rules demonstrated that there is a certain association between the onset of "Type 2 diabetes mellitus" and "Essential (primary) hypertension". Meanwhile, the 40-fold discrepancy in the support of the two rules reflected that "Type 2 diabetes mellitus" was more likely to occur prior to the development of "Essential (primary) hypertension" in the majority of cases. A comparable pattern was also observed among the diseases involved in the remaining 144 rules, demonstrating the directional relationships among the diseases. The greater the disparity between the support values, the more pronounced it is the unidirectional relationship between the two diseases. In terms of the ICD-10 categories, the difference in the support values between the rules "E11 → I10" and "I10 → E11" was the largest, reaching 0.066036233. Additionally, the discrepancy in support values between the rules "I63 → I48" and "I48 → I63" stood at 0.010297326, roughly translating to 2780 instances, making it the second largest, thereby indicating a tendency for "Cerebral infarction" to occur prior to "Atrial fibrillation and flutter". The third largest disparity in support values was observed between the rules "I25 → K29" and "K29 → I25", suggesting a high probability that the progression of disease from "Chronic ischaemic heart disease" to "Gastritis and duodenitis" was unidirectional. Contrary to the three pairs of rules mentioned above, the rules "I63 → H81" and "H81 → I63" represented an extreme opposite, with their support values being very similar. The application of the cSPADE algorithm did not clearly discern the occurrence sequence between "Cerebral infarction" and "Disorders of vestibular function". Consequently, it could only be inferred that there is a symbiotic and correlative relationship between these two diseases. The specific nature of their relationship requires further analyses and validations.

The characteristics of time intervals between illness onset or medical visits

It is well-documented that the cSPADE algorithm controls the time gap between consecutive events in the generated rules through the adjustment of the "mingap" and "maxgap" hyperparameters. We leveraged this feature by setting a variety of "mingap" and "maxgap" values while keeping other hyperparameters constant. This approach allowed us to repeatedly run the algorithm on the same dataset to obtain the time intervals between two instances of illness or healthcare visits, and observe the distribution patterns of these intervals.

We categorized the time differences between consecutive medical visits into 12 windows, spanning from "1 week" to "10 years", and set the hyperparameters to iteratively execute the algorithm for 12 rounds on the dataset according to these intervals. Of the 212 rules, only 154 achieved support greater than 0.001 in at least one of the windows (see Additional file 3). The remaining 58 rules were excluded because their support was below 0.001 in all of the windows. This general deficiency in support was a result of the number of patients conforming to each rule being dispersed across each window, thereby undergoing a "dilution" process.

Among the 154 rules, 100 rules (accounting for 64.94%) achieved their maximum support in the window designated "> 1 year and ≤ 2 years", suggesting that the onset interval for the majority of disease pairs exhibiting comorbid relationships typically ranged from over one year to under two years. The central tendency of the time differences between some diseases was remarkably pronounced. As an illustration, the support level of the rule "C22 → K74" exceeded 0.001 only within the temporal window of "> 6 months and ≤ 1 year" (302 individuals who fell within this stage). The total number of patients who conformed to this rule was 653, indicating that 46.23% of patients would re-consult due to "Fibrosis and cirrhosis of liver" only during the six-month to one-year window following the onset of "Malignant neoplasm of liver and intrahepatic bile ducts". Conversely, the central tendency of the time intervals between diseases was not pronounced in some patterns. Instead, they were evenly distributed across each temporal window. For instance, the support for the rule "E11 → I10" consistently exceeded 0.001 in all temporal windows, indicating

that "Essential (primary) hypertension" can occur at any time within the 10-year period following the onset of "Type 2 diabetes mellitus", despite the fact that the maximum support was achieved in the window designated "> 1 year and ≤ 2 years".

The relationships among diseases were classified into long-term and short-term categories. Nevertheless, in this study, clearly defined short-term or long-term relationships were not frequently observed. Initially, with regard to short-term relationships, five rules involving a time difference of "≤ 2 weeks" between two consultations were identified: "E11 → I10", "I10 → I50", "I10 → I69", "I25 → I20", and "N18 → I10". However, the support for these five rules also exceeded 0.001 across other time intervals, rather than being restricted to the period of "≤ 2 weeks". Indeed, only the support levels of the rules "I63 → G81", "E11 → G81" and "I61 → G81" exceeded 0.001 exclusively during the time window "> 2 weeks and ≤ 2 months", indicating that these rules represent relative short-term relationships. Upon examining the long-term relationships among diseases, the support for 45 rules, including "I10 → K29", surpassed 0.001 during the "> 5 years" window. Nevertheless, these rules also demonstrated support greater than 0.001 across other temporal windows. Hence, more precisely, these rules did not accurately depict long-term relationships between diseases. On the other hand, the support levels of the rules "H25 → I63", "H81 → E11", and "M47 → I20" exceeded 0.001 only in the temporal window of "> 3 years and ≤ 5 years". These rules can be considered representative of the long-term relationships between diseases.

The gender disparities in the rules

To compare the differences in support levels of the 212 rules mentioned above based on gender, the total dataset was partitioned into two subsets based on patients' gender. Subsequently, the cSPADE algorithm was executed separately on the male and female datasets. Overall, the number of male and female patients with multiple medical visits was roughly equivalent. Hence, when employing the algorithm to mine the respective rules for each gender, the support levels of the corresponding rules obtained with the same hyperparameters could be directly compared. Additionally, since the male and female datasets each comprised half of the total dataset, to ensure consistency with the previous sections, the minimum support hyperparameter of the algorithm was set at 0.002. This ensured that a rule would be extracted only if at least 270 patients conformed to it, mirroring the criteria used when mining the entire dataset.

A total of 207 rules exhibited a support level exceeding 0.002 in at least one of the datasets, either the male dataset or the female dataset. The rules "J18 → C34", "I20 → C34", "I63 → S72", "K85 → K80" and "I20 → H81" were excluded by the algorithm because they failed to reach a support level of 0.002 in either the male dataset or the female dataset. A total of 166 rules achieved a support level exceeding 0.002 in both datasets (see Additional file 4). Among these, 146 showed higher support in the male dataset relative to the female dataset, accounting for 87.95%. This prevalence suggested that men were more susceptible to comorbidities or were more likely to seek medical attention. In particular, "C22 → B18" experienced a substantial increase in support in the male dataset compared to the female dataset, with an increase of 359.60%. This finding indicated that males, relative to females, were more likely to develop or seek medical attention specifically regarding "Chronic viral hepatitis", following a diagnosis of "Malignant neoplasm of liver and intrahepatic bile ducts". In contrast, "M81 → I10" experienced the most significant decrease in support in the male dataset relative to the female dataset, with a reduction of 78.67%. This finding implied that 1330 female patients who were diagnosed with "Osteoporosis without pathological fracture" subsequently developed "Essential (primary) hypertension", whereas only 294 male patients experienced the same condition. Moreover, an intriguing observation was made regarding the rules "I20 → G45", "K80 → I63", "K29 → E11" and "K29 → I20" that demonstrated higher support in the female dataset relative to the male dataset. However, their converse rules, namely "G45 → I20", "I63 → K80", "E11 → K29" and "I20 → K29", exhibited higher support in the male dataset relative to the female dataset. It could indicate that the sequence of disease occurrence is related to gender.

The remaining 41 rules demonstrated a support level exceeding 0.002 exclusively in either the male dataset or the female dataset (see Additional file 5). Among these, 30 rules displayed a support level exceeding 0.002 only in the male dataset, while 11 rules achieved a support level above 0.002 exclusively in the female dataset. Particularly, "N40 → I10" attained the highest support level in the male dataset, while "I10 → M81" achieved the highest support level in the female dataset. No instances of the exchange of antecedent and consequent diseases across genders were identified, as described earlier in the text.

Conclusion

The cSPADE algorithm was employed to explore the sequential patterns of disorders within an EMR diagnosis database. The algorithm further investigated the impact of time intervals and gender on comorbidity or multimorbidity by comparing differences in support level, highlighting that the cSPADE algorithm operates effectively on EMR datasets.

This study uncovered comorbidity patterns, detailing the sequence of onset and the intervals between the occurrence of two related diseases. We posit that these

patterns could offer significant technical support for disease prevention, etiological investigations, and the development of CDSS.

Discussion

By comparing the sequential pattern rules unearthed in the current study with the association rules mined by the Apriori algorithm [13], several intriguing insights and revelations were derived. The majority of sequential pattern rules exhibited a trend consistent with the association rules. For instance, we previously discovered an association between Parkinson disease (G20) and Cerebral infarction (I63), but we were unaware of their order of occurrence. In the present study, the support for Parkinson disease that occurred before Cerebral infarction was 0.004411552, while the support for Cerebral infarction that occurred before Parkinson disease was 0.001285314. This indicates that Parkinson disease is more likely to manifest before a Cerebral infarction. This study also confirmed that Cerebral infarction is most likely to occur within a window of more than one year and less than two years after the onset of Parkinson disease. These findings allowed us to predict the developmental trend of Cerebral infarctions with greater accuracy. A similar scenario was distinctly observed in association rules such as "Chronic kidney disease (N18) - Essential (primary) hypertension (I10)" and "Cervical disc disorders (M50) - Cerebral infarction (I63)", where the sequential pattern rules corresponding to these associations indicated that the antecedent disease and consequent disease were fixed and non-interchangeable. It further elucidated the sequence of disease occurrence. Another noteworthy observation was that we did not uncover any sequential pattern rules corresponding to association rules such as "Hypotension (I95) - Cerebral infarction (I63)" and "Malignant neoplasm of renal pelvis (C65) - Malignant neoplasm of kidney, except renal pelvis (C64)". We hypothesized that this was attributable to the inherently low support of these patterns in association rule mining, which diminished further when incorporating temporal attributes, thus failing to surpass the minimum support threshold of 0.001 in this research. The veracity of associations among the diseases within these rules necessitates further exploration.

The largest number of rules related to circulatory system diseases and endocrine system diseases was identified in the final results of this research. This was due to the cSPADE algorithm's reliance on support as a criterion for filtering candidate rules, where support is equivalent to the frequency of occurrence of diseases within a rule. In recent years, the prevalence of circulatory system diseases and endocrine system diseases in China has shown an annual upward trend. According to the most recently published *YEARBOOK OF HEALTH STATISTICS IN CHINA* [25], the prevalence of circulatory system diseases among residents aged over 15 years is 251.0‰, ranking first among all body system diseases. In particular, the prevalence of hypertension is 181.4‰. Meanwhile, the prevalence of endocrine system diseases stands at 62.5‰, ranking second in the disease spectrum, with a prevalence of 53.1‰ for diabetes. This aligns closely with the disease frequency ranking observed in our dataset, which can be considered a microcosm of the disease spectrum of China, thereby rendering the conclusions drawn from this research broader in implication.

Aside from diseases of the circulatory and endocrine systems, conditions affecting other bodily systems also contribute to the comorbidity network, a notion that has been progressively substantiated by existing research. Vakhushev [26] posited that cholelithiasis (K80) might represent a manifestation of systemic digestive disorders. His study, which examined 317 patients with cholelithiasis, revealed that 61.8% of these patients had previously suffered from chronic superficial gastritis, 17.4% from chronic atrophic gastritis, and 18.3% from moderate diffuse duodenitis. Moraes [27] similarly endorsed this conclusion. Zhu et al. [28] investigated the pathological manifestations at different stages in 53 patients with chronic liver disease and hepatic fibrosis and found a progressive increase in spleen volume, which may indicate possible alterations in the spleen. Additionally, it is well-established that chronic viral hepatitis (B18) leads to liver cirrhosis (K74) [29], which is also closely associated with splenic diseases (D73) such as hypersplenism [30], and hepatocellular carcinoma (C22) [31], thereby forming a comprehensive network of comorbidities. When discussing the relationship between malignant neoplasms of the breast (C50) and pulmonary neoplasms (C78), Gonçalves [32] discovered that 7.31% of breast cancer patients developed pulmonary tumors either concurrently or at a later stage. Likewise, to evaluate the risk of lung cancer post-pneumonia, a comprehensive tracking and follow-up study was undertaken on 342,609 patients diagnosed with pneumonia (J18) in all Danish hospitals from 1995 to 2011, and a total of 5887 individuals developed lung cancer (C34) [33]. Notably, 1403 of these lung cancer cases occurred more than five years after the initial pneumonia diagnosis, confirming a long-term relationship between pneumonia and lung cancer.

Some sequential patterns lack sufficient epidemiological or even molecular studies to corroborate them. It does not imply that these patterns can be disregarded. On the contrary, they are precisely the allure of this study and serve as a beacon for future investigations into comorbidity research. Besides, a lack of detailed understanding of molecular pathogenic mechanisms does not prevent these sequential patterns from aiding in the development

of auxiliary diagnostic systems, thereby contributing to alleviating patient suffering more quickly.

Comorbidity pattern research shows regional specificity. Several interesting conclusions can be drawn by comparing the comorbidity patterns identified in this study with those from Jensen's study conducted in Denmark in 2014 [10]. In Denmark, prostate hypertrophy (N40) was found to progress through prostate cancer (C61) and obstructive uropathy (N13), ultimately leading to metastatic cancer (C79) and cancer-associated anemia (D63). However, no evidence was found in our study to suggest that prostate hypertrophy progresses to prostate cancer or metastatic cancer. Instead, we found a notable association between prostate hypertrophy and cardiovascular and cerebrovascular conditions such as cerebral infarction (I63), chronic ischemic heart disease (I25), and essential (primary) hypertension (I10), though the sequence of these events remains uncertain. This discrepancy may stem from the relatively high prevalence of prostate cancer in Western countries compared to its lower prevalence in Asian populations [34]. Similar patterns were observed for conditions such as gout (M10) [35], epilepsy (G40) [36], and retinal disease (H36) [37], all of which exhibit lower prevalence rates in Asian populations. These findings suggest that comorbidity patterns are closely associated with the prevalence of any individual disease within the chain of disease progression. Moreover, the prevalence of disease is shaped by factors such as the genetic background of the study population, the level of medical advancement, the effectiveness of prevention strategies, socioeconomic conditions, and environmental factors in the region under investigation. To better understand these dynamics, we recommend conducting large-scale comorbidity studies every 10 to 20 years, with particular emphasis on comparative analyses across regions.

EMR began to achieve widespread adoption in the three hospitals involved in this study around 2010. However, the quality of the early EMR data was relatively poor due to the immaturity of the technology at the time. This was primarily attributed to incomplete database interfaces, resulting in missing sociological information about patients, such as incomplete identification numbers, gender data, or phone numbers. In this research, we identified similar issues in the records of some patients who visited the hospitals as late as 2012 and 2013. While we were able to partially address these deficiencies by cross-referencing redundant information from other systems, such as the Laboratory Information System (LIS) and the Picture Archiving and Communication System (PACS), a very small fraction of the data (< 0.1%) remained irrecoverable. Given the elapsed time, it was no longer feasible to contact the affected patients to rectify the issues. Consequently, we were compelled to exclude these patients from the final analysis. This limitation may have introduced biases into our findings, particularly in the context of long-term relationship patterns between diseases spanning more than five years. The reduced support level for such patterns increases the likelihood that they were excluded by the algorithm, which may explain the smaller number of long-term relationship patterns identified in this study. In contrast, the support for short-term and medium-term relationship patterns remained unaffected.

In addition to the aforementioned constraint, this study identifies two primary limitations. First, although it was a multicentric study and the distribution of principal diseases in the dataset was similar to that in other regions in China, the inherent limitations of the sequential pattern mining algorithms limited the exploration of comorbidity patterns for rare or endemic diseases. It is imperative to organize and validate larger-scale studies while maintaining a focus on localized details. Second, the cSPADE algorithm was incapable of deriving definitive causal relationships. The sequential patterns it mined held more statistical significance and could not exclude the influence of confounding factors such as age. However, this did not hinder these rules from providing valuable clues for future etiological research.

Abbreviations

SPADE	Sequential Pattern Discovery using Equivalence Classes
CDSS	Clinical Decision Support Systems
EMR	Electronic Medical Records
COVID-19	Corona Virus Disease 2019
ICD-10	The International Classification of Diseases, Tenth Revision
WHO	World Health Organization
uslD	Unique sequence identifier
cSPADE	Constrained SPADE

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13690-025-01589-1.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements
Nothing to declare.

Author contributions
H.M. and L.Z. conceptualized the study. H.M. and H.S. developed the methodology and performed the statistical analysis. Q.H., H.Z. and Y.L. screened and evaluated each pattern to ensure alignment with fundamental medical axioms. B.Z. and L.Z. revised the manuscript. All authors read and approved the final manuscript.

Funding
This research was supported by the Medical Science and Technology Innovation Project of the Health Commission of Xuzhou (XWKYHT20220141) and the Project of the Chinese Hospital Reform and Development Institute,

Nanjing University (NDYG2023035). The funders had no role in the design, execution, interpretation, or decision to publish the results of this study.

Data availability
If needed, authors can provide the anonymous dataset.

Declarations

Ethics approval and consent to participate
Since the study did not involve human experimentation and all the information was anonymized, it was exempt from approval by the Ethics Committee of the Affiliated Hospital of Xuzhou Medical University.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Information and Control Engineering, China University of Mining and Technology, No.1 Daxue Road, Xuzhou 221000, Jiangsu, P.R. China
[2]Department of medical records and statistics, the Affiliated Hospital of Xuzhou Medical University, No.99 Huaihai West Road, Xuzhou 221000, Jiangsu, P.R. China
[3]Department of Interventional Radiology, the Affiliated Hospital of Xuzhou Medical University, No.99 Huaihai West Road, Xuzhou 221000, Jiangsu, P.R. China
[4]Medical records and statistics Center, Xuzhou First People's Hospital, No.269 Daxue Road, Xuzhou 221000, Jiangsu, P.R. China
[5]Quality Control Center, Northern Jiangsu People's Hospital, No.98 Nantong West Road, Yangzhou 225000, Jiangsu, P.R. China
[6]Department of Pharmacy, the Affiliated Hospital of Xuzhou Medical University, No.99 Huaihai West Road, Xuzhou 221000, Jiangsu, P.R. China

Received: 31 October 2024 / Accepted: 30 March 2025
Published online: 10 April 2025

References

1. Slivnick J, Lampert BC. Hypertension and heart failure. Heart Fail Clin. 2019;15(4):531–41.
2. Georgianos PI, Agarwal R. Hypertension in chronic kidney disease-treatment standard 2023. Nephrol Dial Transpl. 2023;38(12):2694–703.
3. Kato H, Mitani Y, Goda T, Yamaue H. Concomitant gallbladder agenesis with methimazole embryopathy. Am J Case Rep. 2020;21:e926310.
4. Shojaeifard M, Saedi S, Alizadeh Ghavidel A, Karimlu MR, Kasaei M, Reza Pouraliakbar H, et al. Concomitant cardiac and hepatic hemangiomas. Echocardiography. 2020;37(3):462–4.
5. Sunwoo BY, Raphelson JR, Malhotra A. Chronic obstructive pulmonary disease and obstructive sleep apnea overlap: who to treat and how? Expert Rev Respir Med. 2024;18(7):527–37.
6. Shin YS, Soni KK, Lee DY, Kam SC. The relationship between depression, anxiety and lower urinary tract symptoms in men. Prostate Int. 2024;12(2):86–9.
7. Gollapudi M, Thomas A, Yogarajah A, Ospina D, Daher JC, Rahman A, et al. Understanding the interplay between premenstrual dysphoric disorder (PMDD) and female sexual dysfunction (FSD). Cureus. 2024;16(6):e62788.
8. Cai Y, Zhou S, Fan S, Yang Y, Tian K, Luo L, et al. The Multimorbidity association of metabolic syndrome and depression on type 2 diabetes: a general population cohort study in Southwest China. Front Endocrinol (Lausanne). 2024;15:1399859.
9. Keita M, Seck M, Diallo AB, Touré SA, Bousso ES, Gueye SM et al. Morbidity and Mortality Associated with COVID-19 and Acute Chest Syndrome in Sickle Cell Disease Patients. Hemoglobin. 2024:1–7.
10. 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. Nat Commun. 2014;5(1):4022.
11. Fei Z, Ryeznik Y, Sverdlov O, Tan CW, Wong WK. An overview of healthcare data analytics with applications to the COVID-19 pandemic. IEEE Trans Big Data. 2022;8(6):1463–80.
12. Hang C-N, Tsai Y-Z, Yu P-D, Chen J, Tan C-W. Privacy-Enhancing digital contact tracing with machine learning for pandemic response: A comprehensive review. Big Data Cogn Comput. 2023;7(2):108.
13. Ma H, Ding J, Liu M, Liu Y. Connections between various disorders: combination pattern mining using apriori algorithm based on diagnosis information from electronic medical records. Biomed Res Int. 2022;2022:2199317.
14. Nawaz MS, Fournier-Viger P, Nawaz S, Zhu H, Yun U. SPM4GAC: SPM based approach for genome analysis and classification of macromolecules. Int J Biol Macromol. 2024;266(Pt 2):130984.
15. Wu J, Liu D, Guo Z, Xu Q, Wu Y, TacticFlow. Visual analytics of Ever-Changing tactics in racket sports. IEEE Trans Vis Comput Graph. 2022;28(1):835–45.
16. Mluba HS, Atif O, Lee J, Park D, Chung Y. Pattern Mining-Based pig behavior analysis for health and welfare monitoring. Sens (Basel). 2024;24(7):2185.
17. Santhiran R, Varathan KD, Chiam YK. Feature extraction from customer reviews using enhanced rules. PeerJ Comput Sci. 2024;10:e1821.
18. Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. Mach Learn. 2001;42(1–2):31–60.
19. Yang Q, Luo T, Zhang W, Zhong X, He P, Zheng H. Data-driven treatment pathways mining for early breast cancer using cSPADE algorithm and system clustering. Int J Health Plann Manage. 2022;37(5):2569–84.
20. Campbell EA, Qian T, Miller JM, Bass EJ, Masino AJ. Identification of Temporal condition patterns associated with pediatric obesity incidence using sequence mining and big data. Int J Obes (Lond). 2020;44(8):1753–65.
21. Wang Y, Hou W, Wang F. Mining co-occurrence and sequence patterns from cancer diagnoses in new York state. PLoS ONE. 2018;13(4):e0194407.
22. Verma N, Mehta N. Sequential pattern mining: A comparison between GSP, SPADE and prefix SPAN. Int J Eng Dev Res. 2014;2(3):3016–36.
23. Pei J, Han JW, Mortazavi-Asl B, Wang JY, Pinto H, Chen QM, et al. Mining sequential patterns by pattern-growth: the prefixspan approach. IEEE Trans Knowl Data Eng. 2004;16(11):1424–40.
24. Y PSJG. Rare diseases, orphan drugs, and their regulation in Asia: current status and. Intractable Rare Dis Res. 2012;1(1):3–9.
25. 2022 National Health Statistics Yearbook of China Beijing: Statistical Information Center of National Health Commission. 2023 [Available from: http://www.nhc.gov.cn/mohwsbwstjxxzx/tjtjnj/202305/6ef68aac6bd14c1eb9375e01a0faa1fb.shtml. (accessed 8 Jun 2024).
26. Vakhrushev YM, Gorbunov AY, Tronina DV, Suchkova EV, Lyapina MV, Khokhlacheva NA. [Cholelithiasis as a possible manifestation of systemic digestive diseases]. Ter Arkh. 2015;87(2):54–8.
27. Moraes AB, Treistman N, Studart MC, Chagas VLA, Brabo EP, Vieira Neto L. Gastrinoma of cystic duct: A rare association with multiple endocrine neoplasia type 1. J Clin Med Res. 2018;10(11):843–7.
28. Zhu TT, Li ZX, Yuan J, Huang K, Chen GF, Guo RF, et al. [Characteristics of liver volume and pathological changes with different stages of liver fibrosis in chronic liver disease]. Zhonghua Gan Zang Bing Za Zhi. 2024;32(6):517–24.
29. Carrier P, Debette-Gratien M, Jacques J, Loustaud-Ratti V. Cirrhotic patients and older people. World J Hepatol. 2019;11(9):663–77.
30. Colombo C, Lanfranchi C, Tosetti G, Corti F, Primignani M. Management of liver disease and portal hypertension in cystic fibrosis: a review. Expert Rev Respir Med. 2024;18(5):269–81.
31. Perisetti A, Goyal H, Yendala R, Thandassery RB, Giorgakis E. Non-cirrhotic hepatocellular carcinoma in chronic viral hepatitis: current insights and advancements. World J Gastroenterol. 2021;27(24):3466–82.
32. Gonçalves E, Fontes F, Rodrigues JR, Calisto R, Bento MJ, Lunet N, et al. The contribution of second primary cancers to the mortality of patients with a first primary breast cancer. Breast Cancer Res Treat. 2024;207(2):323–30.
33. Søgaard KK, Farkas DK, Pedersen L, Weiss NS, Thomsen RW, Sørensen HT. Pneumonia and the incidence of cancer: a Danish nationwide cohort study. J Intern Med. 2015;277(4):429–38.
34. Rawla P. Epidemiology of prostate cancer. World J Oncol. 2019;10(2):63–89.
35. Punjwani S, Jani C, Liu W, Kakoullis L, Salciccioli I, Al Omari O, et al. Burden of gout among different WHO regions, 1990–2019: estimates from the global burden of disease study. Sci Rep. 2024;14(1):15953.
36. Liu W, Xu Y, Lin Y, Wang L, Zhou M, Yin P, et al. Burden of epilepsy in China and its provinces, 1990 to 2019: findings from the global burden of disease study 2019. Chin Med J (Engl). 2023;136(3):305–12.
37. Chiang PP, Lamoureux EL, Cheung CY, Sabanayagam C, Wong W, Tai ES, et al. Racial differences in the prevalence of diabetes but not diabetic

retinopathy in a multi-ethnic Asian population. Invest Ophthalmol Vis Sci. 2011;52(10):7586–92.

Publisher's note