Journal of the American Medical Informatics Association, 27(4), 2020, 558–566
doi: 10.1093/jamia/ocaa005
Advance Access Publication Date: 12 February 2020
Research and Applications

Research and Applications

Temporal condition pattern mining in large, sparse electronic health record data: A case study in characterizing pediatric asthma

Elizabeth A. Campbell,[1,2] Ellen J. Bass,[1,3] and Aaron J. Masino[2,4]

[1]Department of Information Science, College of Computing & Informatics, Drexel University, Philadelphia, Pennsylvania, USA, [2]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA, [3]Department of Health Systems and Sciences Research, College of Nursing & Health Professions, Philadelphia, Pennsylvania, USA, and [4]Department of Anesthesiology and Critical Care, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author: Aaron J. Masino, PhD, 2716 South Street, Room 15362, Philadelphia, PA 19146; masinoa@email.chop.edu

Received 19 September 2019; Revised 20 December 2019; Editorial Decision 7 January 2020; Accepted 12 January 2020

ABSTRACT

Objective: This study introduces a temporal condition pattern mining methodology to address the sparse nature of coded condition concept utilization in electronic health record data. As a validation study, we applied this method to reveal condition patterns surrounding an initial diagnosis of pediatric asthma.

Materials and Methods: The SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm was used to identify common temporal condition patterns surrounding the initial diagnosis of pediatric asthma in a study population of 71 824 patients from the Children's Hospital of Philadelphia. SPADE was applied to a dataset with diagnoses coded using International Classification of Diseases (ICD) concepts and separately to a dataset with the ICD codes mapped to their corresponding expanded diagnostic clusters (EDCs). Common temporal condition patterns surrounding the initial diagnosis of pediatric asthma ascertained by SPADE from both the ICD and EDC datasets were compared.

Results: SPADE identified 36 unique diagnoses in the mapped EDC dataset, whereas only 19 were recognized in the ICD dataset. Temporal trends in condition diagnoses ascertained from the EDC data were not discoverable in the ICD dataset.

Discussion: Mining frequent temporal condition patterns from large electronic health record datasets may reveal previously unknown associations between diagnoses that could inform future research into causation or other relationships. Mapping sparsely coded medical concepts into homogenous groups was essential to discovering potentially useful information from our dataset.

Conclusions: We expect that the presented methodology is applicable to the study of diagnostic trajectories for other clinical conditions and can be extended to study temporal patterns of other coded medical concepts such as medications and procedures.

Key words: data science, data mining, asthma, electronic health records

INTRODUCTION

Data mining refers to the computational process of automated information extraction from large datasets to facilitate discovery of novel insights.[1] Pattern mining is a fundamental data mining task.[2] Important pattern types include subsequences of sequentially ordered items or events that occur frequently in the dataset.[3] For example, a collection of temporally ordered event sequences may contain ordered event subsets that are frequent. Such temporal patterns may yield valuable insights on associations or causal relationships among variables in a dataset, and help to predict future events.

Temporal pattern mining of electronic health record (EHR) data has the potential to uncover previously unknown relationships among comorbidities (conditions occurring together) and condition trajectories (conditions diagnosed in a temporal order), which can complement clinical knowledge and traditional medical research methods. Diagnoses found to commonly occur before the onset of a condition of interest may inform clinicians of patient risk for developing that condition. Similarly, diagnoses found to occur commonly after the onset of a condition of interest may inform care providers of future risk for other disorders.

While longitudinal EHR data may inform healthcare research and policy, data mining methods must be selected carefully based on the characteristics of the EHR data and the outcome of interest.[4] Sequential pattern mining methods can be broadly categorized into 2 classes: the Apriori-based candidate generation method and the pattern growth method.[5] The Apriori approach is based on the Apriori property, which posits that a given sequence is necessarily infrequent if it contains a smaller sequence that is infrequent, and that a subsequence of a frequent sequence is also frequent.[6,7] Apriori-based algorithms may use a horizontal database format (eg, the generalized sequential patterns algorithm),[8] a vertical database format (eg, SPADE [Sequential PAttern Discovery using Equivalence classes]), or Apriori-based candidate generation and pruning using depth-first traversal (eg, the SPAM [Sequential PAttern Mining] algorithm).[9,10] Pattern growth algorithms, such as PrefixSpan, perform database projection and use database scans to count item support (all patterns present above a threshold percentage) but do not generate candidates.[11]

Although the application of temporal pattern mining techniques to EHR data remains a relatively new and unexplored approach,[12] there have been some important initial studies that have demonstrated the methods' effectiveness. For example, Perer et al[13] used the SPAM algorithm to mine EHR data to identify sequential medical event patterns among hyperlipidemic patients and studied their association with outcomes. The study identified sequences that confirmed known associations and revealed novel information. For example, increased use of certain medications, such as fluoroquinolones, in hyperlipidemic patients with hypertension may raise low-density lipoprotein levels. Gotz et al[14] also used the SPAM algorithm for temporal pattern mining of clinical events in retrospective EHR data. Chen et al[15] used the PrefixSpan algorithm to analyze combinations of chronic diseases and specific orders of chronic disease transition from a large medical database. Wright et al[16] utilized the SPADE algorithm to identify temporal relationships among diabetes medication prescription trajectories and to predict the future medication prescriptions.

Data heterogeneity and sparsity are major challenges to temporal pattern mining of EHR data.[17] These challenges are particularly salient when working with diagnostic codes. For example, although a typical EHR dataset will contain thousands of distinct International Classification of Diseases (ICD) codes,[17] a small fraction of ICD codes will account for most diagnoses. Previous research indicates that this Pareto principle[18] phenomenon presents a significant challenge for pattern mining. Chen et al[15] noted the complexities of mining disease transition patterns using ICD codes due the volume of codes and the sparse use of medically similar codes. Boytcheva et al[19] observed similar challenges and highlighted analyzing classes of similarly grouped ICD codes as an important area of future work.

Diagnostic schemes that group diagnosis codes into medically homogenous clusters are one approach to reduce sparsity in an EHR dataset, which may improve pattern mining results. However, the use of such clustering methods in combination with sequential pattern mining in EHR data is limited.[13,14] There is also limited comparison of mining results when utilizing diagnostic schemes compared with using standard codes (ie, ICD codes).[17] To address this, our study utilizes expanded diagnostic clusters (EDCs)[20] from the Adjusted Clinical Group (ACG) System to address sparsity and makes a direct comparison to results using the original ICD-based diagnoses.[21] We selected EDCs because they are linked to both ICD–Ninth Revision–Clinical Modification (ICD-9-CM) and –ICD–Tenth Revision–Clinical Modification (ICD-10-CM), and cover all diagnosis codes, thus providing comprehensive coverage.

We present a sequential pattern mining approach that combines the SPADE pattern mining algorithm with the use of EDC groupings to identify temporal condition patterns in a large, sparse EHR dataset with a case study of pediatric asthma. Pediatric asthma is one of the most common childhood chronic conditions, has numerous comorbidities, and is a socially significant health condition that disproportionately impacts low-income and minority children in the United States.[22,23]

Our objective is to identify temporal patterns surrounding an incident diagnosis of a given condition (eg, asthma). Toward this end, we sought to develop a generalizable framework to identify temporal patterns utilizing sparse EHR data and a sequential pattern mining algorithm. We also consider the impact of collapsing diagnostic codes into clinically similar groups on temporal pattern recognition. We present our data extraction, data transformation, and knowledge discovery processes. We describe the implementation of the SPADE algorithm on EHR data to ascertain condition patterns associated with pediatric asthma using both ICD and EDC coded conditions, and compare output from both datasets. We discuss the evaluation and interpretation of SPADE output, as well as the successes and limitations of our approach. Our study contributes a methodological framework for extracting and organizing EHR data into temporal sequences, and an approach to analyze temporal patterns in EHR data for associations between condition trajectories and an outcome of interest (eg, pediatric asthma).

MATERIALS AND METHODS

The institutional review board at the Children's Hospital of Philadelphia (CHOP) approved this research study through institutional review board protocol number 16-012822 and waived the requirement for consent. The primary methodological steps in the study, from data extraction through algorithm implementation, are described in Figure 1.

Setting

Study data were obtained from the Pediatric Big Data (PBD) resource maintained at CHOP. The PBD resource includes EHR data

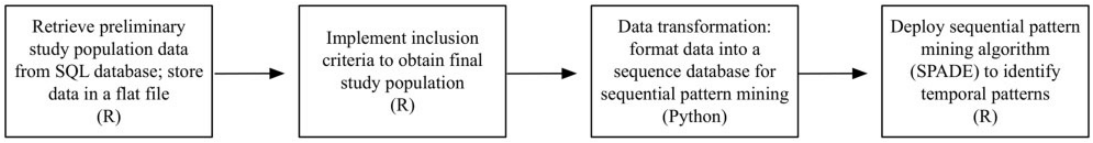

Figure 1. The primary methodological steps involved in the study. SPADE: Sequential PAttern Discovery using Equivalence classes.

from CHOP, its primary care network, and specialty care and surgical centers. EHR information for conditions recorded during each patient visit was derived from ICD-9-CM and ICD-10-CM diagnostic information.[24,25] Data were extracted by nonstudy staff personnel and anonymized to remove all personal health information before transfer. The PBD resource contains data for visits through October 2017.

Inclusion criteria

Patient data were included for individuals with at least 2 clinical encounters on or after January 1, 2005, in which an ICD code for an asthma diagnosis or exacerbation (ICD-9-CM codes 493.* and ICD-10-CM codes J45.*, where * is any child code) (see Supplementary Table 1) was recorded and the encounters were face to face (inpatient stays, ambulatory visits, or emergency department visits). The requirement for 2 records with an asthma diagnosis was meant to exclude transient asthmas diagnoses.

Patients must also have had 2 clinical encounters of any type (not necessarily face to face, such as a pharmacy visit) with a recorded ICD condition concept that was not an asthma diagnosis or exacerbation code (these encounters must have occurred on or after January 1, 2000). This requirement ensures that the cohort does not include individuals with only asthma diagnoses to enable identification of clinically informative temporal condition patterns surrounding an initial diagnosis of pediatric asthma. ICD codes may be classified as clinical or nonclinical observations, per Observational Health Data Sciences and Informatics condition domain standards. For example, an anemia screening is a nonclinical observation, whereas an anemia diagnosis is a clinical observation.[26] In this study, we excluded encounters without any clinical observations.

We extracted the available clinical diagnosis observations for all patients that met the inclusion criteria in the PBD database. This yielded 7 412 524 clinical observations for 94 343 patients. The encounters where asthma conditions were recorded corresponded to 1 135 006 clinical observations. For each individual, we identified the first visit with a recorded asthma diagnosis (the index visit), the encounter immediately before the index visit (the preindex visit, if one exists), and the encounter immediately after the index visit (the postindex visit, if one exists). This yielded a dataset containing 578 839 clinical observations for 94 343 patients. To ensure the ability to analyze patients' clinical state before the initial asthma diagnosis, patients without a preindex visit were excluded. The final dataset contained 473 607 clinical observations for 71 824 patients.

Data extraction and transformation

R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria)[27] and Python 3 (Python Software Foundation, Wilmington, DE)[28] were used for the data analysis. Computations were performed on a MacBook Pro running MacOS version 10.12.6 and with 8 GB of RAM (Apple Inc, Cupertino, CA).

After extracting condition occurrence data from the PBD resource, we mapped ICD-9-CM and ICD-10-CM clinical condition concepts into medically homogenous classes using EDCs[20] from the ACG System.[21]

Temporal order identification

The SPADE algorithm analyzes data in a "vertical id-list" database format, in which sequences comprise a list of objects in order of occurrence along with timestamps.[29] To apply the SPADE algorithm, we reorganized PBD data from their normalized relational form (information for a single patient stored in different tables within the database) into row entries, in which each row contained a patient identifier, a visit timing class indicator (preindex, index, or postindex) and all clinical observations associated with that visit. Visit date information was no longer relevant, as the visit timing class variable captured the temporal order information necessary to indicate sequential order for subsequent SPADE analysis. Figure 2A presents a sample sequence database of clinical information for 3 patients, formatted to the SPADE algorithm's specifications for pattern mining.

Sequence mining

Our primary study objective required identification of temporal condition patterns in a large, sparse EHR dataset. Prior research has demonstrated that SPADE performs well on data with such characteristics and demonstrates runtime efficiency and low memory usage.[30] For these reasons, we selected SPADE as our mining algorithm for this study.

SPADE first identifies individual items (eg, a singular diagnosis in a specific timing class) above a specified support level (eg, the proportion of patients with an identified condition pattern), and then builds more complex sequences (multiple diagnoses across different timing classes) at the given support level. An inherent but provable assumption is that a complex sequence present above a given support level is comprised of individual items that also occur above the support level. The algorithm uses efficient lattice search techniques and simple join operations to find frequent patterns among smaller subproblems obtained from dividing the original problem.[29] The SPADE algorithm is efficient for identifying subsequence patterns in large databases of sequential data, and has been successfully used across numerous domains including internet user behaviors[31] and food purchasing patterns.[32]

We applied the SPADE algorithm as implemented in the R arules package[33] to discover common subsequences in clinical diagnoses among patients in the study population. Figure 2B illustrates the sequential patterns identified by the SPADE algorithm for 3 hypothetical patients with clinical history shown in Figure 2A. Figure 3 illustrates the possible combinations of temporal diagnoses that SPADE detects. It is important to note that frequent patterns may not contain diagnoses from all 3 timing classes. As illustrated in Figure 3, condition patterns may comprise diagnoses that occurred in the pre- and postindex visits (Figure 3A), index and postindex visits (Figure 3B), preindex and index visits (Figure 3C), and across all 3 timing classes (Figure 3D). Additionally, SPADE identifies singular

(a)

Patient	Timing Class	Clinical Conditions
A	1	1-EAR08
A	2	2-MUS02, 2-MUS08
A	3	3-MUS08
B	1	1-EAR08
B	2	2-GAS03
C	1	1-EAR08, 1-EYE07
C	2	2-ALL04

(b)

Sequence	Support
<{1-EAR08}>	1
<{1-EYE07}>	0.33
<{2-ALL04}>	0.33
<{2-GAS03}>	0.33
<{2-MUS02}>	0.33
<{2-MUS08}>	0.33
<{3-MUS08}>	0.33
<{2-MUS02}, {3-MUS08}>	0.33
<{2-MUS08}, {3-MUS08}>	0.33
<{2-MUS02,2-MUS08}, {3-MUS08}>	0.33
<{1-EAR08}, {2-MUS02,2-MUS08}, {3-MUS08}>	0.33
<{1-EAR08}, {2-MUS08}, {3-MUS08}>	0.33
<{1-EAR08}, {2-MUS02}, {3-MUS08}>	0.33
<{1-EAR08}, {2-MUS08}>	0.33
<{2-MUS02,2-MUS08}>	0.33
<{1-EAR08}, {2-MUS02,2-MUS08}>	0.33
<{1-EAR08}, {2-MUS02}>	0.33
<{1-EAR08}, {2-GAS03}>	0.33
<{1-EAR08}, {2-ALL04}>	0.33
<{1-EYE07}, {2-ALL04}>	0.33
<{1-EAR08,1-EYE07}, {2-ALL04}>	0.33
<{1-EAR08,1-EYE07}>	0.33

Figure 2. (A) A sample sequence database of clinical information (ie, expanded diagnostic cluster groupings for clinical diagnoses) for 3 patients, formatted for the SPADE (Sequential PAttern Discovery using Equivalence classes) algorithm. (B) The output of frequent sequential patterns that SPADE uncovers from the clinical information of the 3 patients in panel A. The expanded diagnostic cluster codes correspond to the following diagnostic categories: ALL04: asthma, without status asthmaticus; EAR08: deafness, hearing loss; EYE07: conjunctivitis, keratitis; GAS03: constipation; MUS02: acute sprains and strains; MUS08: fractures and dislocations/digits only.

temporal diagnoses that occur among patients at the specified minimum support level (eg, a diagnosis of the EDC code ALL04 [asthma] in the index visit).

We wanted to identify as many common condition patterns as possible while ensuring that the patterns were present among a meaningful number of patients; therefore, we selected a support level of 0.01 to identify all condition patterns that were present in 1% or more of patients. To characterize the impact of grouping clinically similar conditions, we deployed SPADE on both the full ICD dataset and the mapped EDC dataset. For ease of interpretability and results comparison, the common condition patterns identified by SPADE in the ICD dataset were then mapped from ICD codes to EDC codes. While the support values would not change for the common condition patterns identified from the ICD dataset once they have been mapped to EDC codes, it is possible for 2 or more patterns to appear to have the same temporal and diagnostic information, since more than 1 ICD code is mapped to a single EDC code.

We analyzed our results for clinical relevance, and compared the prevalence of condition patterns by support level. We identified sequences with the highest support level, which have the strongest potential associations with pediatric asthma. As support is invariant to timing class, we also examined trends of condition diagnoses relative to the visit timing class.

RESULTS

Clinical term mapping

There were 7072 unique ICD codes represented in the dataset, which mapped to 267 EDC codes.

SPADE analysis

When deployed on the mapped EDC dataset, SPADE's runtime was 0.72 seconds, and it identified 439 sequences with a support level of 0.01 or higher. Support for these sequences ranged from 0.01 to 0.94. The mean and median level of support were 0.03 and 0.019, respectively. When deployed on the ICD dataset, SPADE's runtime was 0.49 seconds, and it identified 203 sequences with a support level of 0.01 or higher. Support for these sequences ranged from 0.01 to 0.54. The mean and median level of support were 0.03 and 0.018, respectively.

Table 1 illustrates the top 20 condition patterns with the highest level of support (0.09-0.94) in the study population identified in the EDC dataset. Among the most prevalent sequences, 5 distinct conditions were present: asthma without status asthmaticus, respiratory signs and symptoms, acute upper respiratory tract infection, allergic rhinitis, and otitis media.

Table 2 illustrates the conditions present in at least 1 of 439 sequences identified by SPADE in the EDC dataset. There were 36 unique EDC codes represented among the identified sequences. The majority of conditions were present in all 3 visit timing classes. However, there were 5 EDC codes that were identified in the preindex visit only: exanthems, nausea/vomiting, dermatophytosis, nonfungal infections of skin and subcutaneous tissue, and allergic reactions. No conditions were present exclusively in the postindex visit.

Table 3 presents the top 20 condition patterns with the highest level of support (0.06-0.54) identified by SPADE in the full ICD dataset (the ICD concepts in the identified sequences were mapped to EDC codes for comparison, with results in Table 1).

Among the most prevalent sequences, 6 distinct conditions were present: asthma without status asthmaticus, respiratory signs and symptoms, acute upper respiratory tract infection, allergic rhinitis, cough, and otitis media. Table 4 shows the unique conditions present in the 203 sequences identified by SPADE in the ICD dataset (again mapped to EDC codes for ease of comparison). There were 19 unique EDC codes present. The majority of conditions occurred

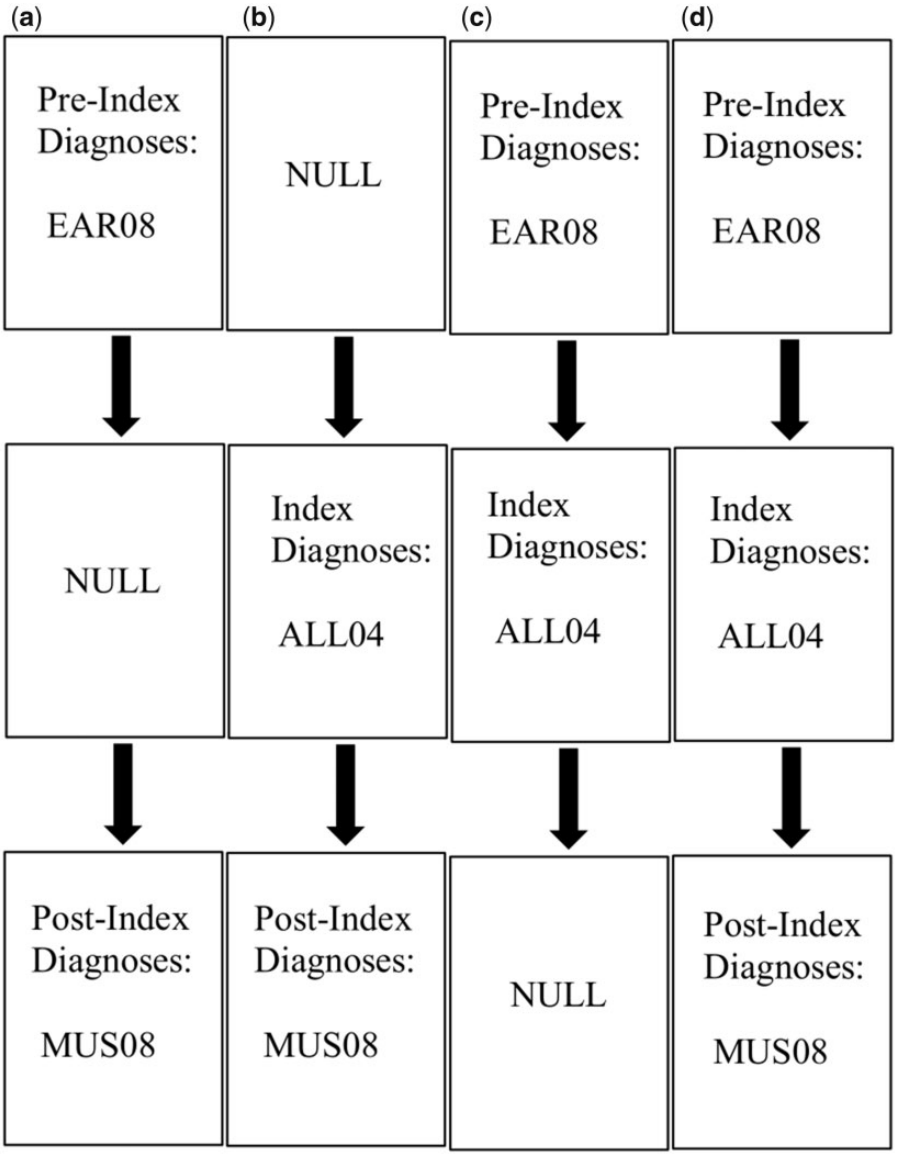

Figure 3. Sample temporal diagnostic sequences discoverable by SPADE (Sequential PAttern Discovery using Equivalence classes). (A) A sequence that includes a clinical diagnosis in the preindex visit and the postindex visit. (B) A sequence that includes a clinical diagnosis in the index visit and the postindex visit. (C) A sequence that includes a clinical diagnosis in the preindex visit and the index visit. (D) A sequence that includes a clinical diagnosis in all 3 timing classes. Clinical diagnoses observed in a single timing class may also be considered common sequences.

in all 3 visit timing classes. Only gastroenteritis was diagnosed solely in the preindex visit. No conditions were present exclusively in the postindex visit.

Among the top 20 most prevalent patterns observed in the EDC dataset, SPADE detected 3 condition patterns with diagnoses observed in the preindex and index visits and 3 condition patterns with diagnoses observed in the index and postindex visits. Three singular patterns of diagnoses in the preindex visit, 7 in the index visit, and 4 in the postindex visit were observed. Among the top 20 most prevalent patterns in the ICD dataset, SPADE detected 1 condition pattern with diagnoses observed in the preindex and index visits and 1 condition pattern with diagnoses observed in the index and postindex visits. Four singular patterns of diagnoses in the preindex visit, 10 in the index visit, and 4 in the postindex visit were identified.

DISCUSSION

Methodological contributions

We presented a novel methodology to mine EHR data for temporal condition patterns associated with pediatric asthma. Our approach included extraction of relevant information from a large EHR database, temporal data organization, selection of a sequential pattern mining algorithm, and grouping of clinically similar codes to address data sparsity. The methodology described in this study can be used to study temporal condition patterns within large volumes of EHR data regardless of the condition of interest. We found the SPADE algorithm to be a time-efficient approach to mining a large, sparse dataset for frequent longitudinal clinical condition patterns surrounding initial diagnosis of a selected condition of interest (pediatric asthma in this study).

Table 1. Top 20 most prevalent conditions patterns surrounding initial diagnosis of pediatric asthma (EDC dataset)

Preindex visit condition(s)	Index visit condition(s)	Postindex visit condition(s)	Support
	Asthma without status asthmaticus		0.942
		Asthma without status asthmaticus	0.436
	Asthma without status asthmaticus	Asthma without status asthmaticus	0.417
Acute upper respiratory tract infection			0.186
Acute upper respiratory tract infection	Asthma without status asthmaticus		0.175
	Allergic rhinitis		0.164
	Allergic rhinitis, asthma without status asthmaticus		0.158
		Acute upper respiratory tract infection	0.145
Respiratory signs and symptoms			0.139
	Acute upper respiratory tract infection		0.139
Otitis media			0.137
	Asthma without status asthmaticus	Acute upper respiratory tract infection	0.136
Respiratory signs and symptoms	Asthma without status asthmaticus		0.134
	Asthma without status asthmaticus, Acute upper respiratory tract infection		0.132
Otitis media	Asthma without status asthmaticus		0.131
		Otitis media	0.120
	Asthma without status asthmaticus	Otitis media	0.116
		Allergic rhinitis	0.102
	Otitis media		0.101
	Asthma without status asthmaticus, Otitis media		0.098

EDC: expanded diagnostic cluster.

Table 2. Conditions observed in prevalent sequences identified by SPADE (EDC dataset), by timing class

All visit classes	Preindex visits	Preindex and index visits	Preindex and postindex visits	Index visits	Index and postindex visits
Acute lower respiratory tract infection	Allergic reactions	Chronic pharyngitis and tonsillitis	Abdominal pain	Seizure disorder	Asthma without status asthmaticus
Acute upper respiratory tract infection	Dermatophytosis	Musculoskeletal disorders, other	Acute sprains and strains		Asthma, with status asthmaticus
Administrative concerns and nonspecific laboratory abnormalities	Exanthems		Contusions and abrasions		
Allergic rhinitis	Nausea, vomiting		Deafness, hearing loss		
Attention-deficit disorder	Nonfungal infections of skin and subcutaneous tissue		Musculoskeletal signs and symptoms		
Conjunctivitis, keratitis			Urinary symptoms		
Constipation					
Cough					
Dermatitis and eczema					
Developmental disorder					
ENT disorders, other					
Failure to thrive					
Gastroenteritis					
Gastroesophageal reflux					
Nonspecific signs and symptoms					
Obesity					
Otitis media					
Respiratory signs and symptoms					
Sinusitis					
Viral syndromes					

EDC: expanded diagnostic cluster; ENT: ear, nose, and throat; SPADE: Sequential PAttern Discovery using Equivalence classes.

In recent years, the generation and use of large datasets derived from nontraditional data sources (eg, EHRs) in the healthcare sector has increased.[34,35] The generation of such data enable the extraction of useful and potentially previously unknown insights from massive datasets, but only with the proper computational methods to convert this data into clinically meaningful information and knowledge.[36,37] Our work provides a case study into a successful aggregation method and examination of condition information from

Table 3. Top 20 most prevalent conditions patterns surrounding initial diagnosis of pediatric asthma (ICD dataset)

Preindex visit condition(s)	Index visit condition(s)	Postindex visit condition(s)	Support
	Asthma without status asthmaticus		0.538
		Asthma without status asthmaticus	0.251
	Asthma without status asthmaticus	Asthma without status asthmaticus	0.183
	Asthma without status asthmaticus		0.149
	Asthma without status asthmaticus		0.137
	Allergic rhinitis		0.125
Respiratory signs and symptoms			0.108
Acute upper respiratory tract infection			0.096
	Acute upper respiratory tract infection		0.088
	Cough		0.081
	Asthma without status asthmaticus		0.079
		Allergic rhinitis	0.077
	Allergic rhinitis, asthma without status asthmaticus		0.075
		Acute upper respiratory tract infection	0.072
Respiratory signs and symptoms	Asthma without status asthmaticus		0.064
		Asthma without status asthmaticus	0.061
Acute upper respiratory tract infection			0.059
	Cough		0.059
Otitis media			0.057
	Asthma without status asthmaticus		0.055

ICD: International Classification of Diseases.

Table 4. Conditions observed in prevalent sequences identified by SPADE (ICD dataset), by timing class

All visit classes	Preindex visits	Preindex and postindex visits	Index visits	Index and postindex visits
Acute lower respiratory tract infection	Gastroenteritis	Deafness, hearing loss	Attention-deficit disorder	Asthma without status asthmaticus
Acute upper respiratory tract infection			Asthma, with status asthmaticus	
Administrative concerns and nonspecific laboratory abnormalities			Obesity	
Allergic rhinitis				
Constipation				
Cough				
Dermatitis and eczema				
Failure to thrive				
Gastroesophageal reflux				
Otitis media				
Respiratory signs and symptoms				
Sinusitis				
Viral syndromes				

ICD: International Classification of Diseases; SPADE: Sequential PAttern Discovery using Equivalence classes.

EHR data that can be directly applied to similar datasets. Our study specifically demonstrated the utility of EDC codes when mining large EHR datasets to reveal additional important patterns.

When deployed on the full dataset of ICD codes, SPADE identified fewer than half the number of common condition patterns than when deployed on the dataset containing EDC codes. SPADE was also able to better detect more complex temporal patterns in diagnoses in the EDC dataset. By using EDC codes, we consolidated clinically similar results, detected common condition patterns at higher levels of support, and found condition patterns at low levels of support that would otherwise not have been identified. The EDC mapping optimized the recognition of potential condition associations (36 unique diagnoses were identified in the mapped EDC dataset, compared with only 19 in the ICD dataset), and was essential to analyzing temporal trends in condition trajectories. Gastroenteritis was the only diagnosis exclusively found in the preindex visit in the ICD dataset. However, in the EDC dataset, gastroenteritis diagnoses were found across all timing classes, and 5 separate conditions were identified as occurring exclusively in the preindex visit. None of these diagnoses were identified by SPADE in the ICD dataset. While our approach utilized the ACG system to aggregate related, but sparsely used clinical concepts, it is important to note that other methods (eg, the Clinical Classifications Software for ICD-9-CM) exist for clustering patient diagnoses into a manageable number of clinically related categories for analysis.[38]

To our knowledge, this is the first study to utilize the SPADE algorithm to study temporal condition patterns surrounding initial diagnosis of pediatric asthma. The patterns identified by SPADE can be regarded as hypotheses for associations between specific conditions and temporal patterns and pediatric asthma that future studies can explore. We found strong associations among allergic rhinitis, respiratory signs and symptoms, acute upper respiratory tract infec-

tion, and otitis media and pediatric asthma. Five conditions (exanthems, nausea/vomiting, dermatophytosis, nonfungal infections of skin and subcutaneous tissue, and allergic reactions) were present exclusively in the preindex visit among common sequences. These conditions may represent signals of future asthma that could alert clinicians. Future research into the reproducibility and potential causal relations of these associations is warranted.

Diagnostic pattern associations

The patterns with the highest support discovered by SPADE in this study align with clinical knowledge of pediatric asthma comorbidities. Allergic rhinitis,[39] sinusitis,[40] eczema, and respiratory infections[41] have previously been shown to be associated with pediatric asthma. As these findings are consistent with prior epidemiological asthma studies, we may speculate that the other condition patterns found in this study may also be clinically relevant and are possible areas for future epidemiological research. Furthermore, this approach may reveal information about disease comorbidities. For example, although it is known that asthmatic children with multiple morbidities have increased asthma symptoms, school absences, and visits to emergency departments,[42] knowledge of pediatric asthma comorbidities remains an understudied topic.[43] Our methodology represents a scalable approach to study comorbidity patterns that may be employed to explore pediatric asthma and other critical health conditions.

The methods described in this study differ from other epidemiological approaches, such as those described in Beck et al[44,45] which focused on the predictive power of disease trajectories uncovered from EHR data to assess relative risk of sepsis mortality. Rather than studying the predictive power of a particular sequence, our work serves to describe an approach to analyze complex datasets to uncover patterns and generate hypotheses for future research. While we focused on pediatric asthma, the methodology is not dependent on the conditions analyzed and can be utilized to uncover temporal patterns surrounding any health outcome (eg, diabetes or obesity) or medical event of interest (eg, medication usage). Such an approach may potentially uncover previously unknown associations that have clinical and public health utility for researchers studying asthma and other health outcomes.

Limitations

We utilized a large and complex dataset, which required extensive personnel time and resources to obtain and analyze. As EHRs are primarily designed for clinical care and billing purposes, there are challenges to utilizing the data for clinical and translational research.[46] Although groups such as Observational Health Data Sciences and Informatics have recommended that EHR data be organized temporally,[47] currently most EHRs are stored in databases with various data points for a single patient stored in different tables across schemas within the database.[48] As EHR data are typically not temporally organized in a singular table, relevant data must be identified across the database and transformed into temporal sequences retrospectively for analysis to mine temporal patterns. It is important to note that our dataset was developed from a secondary-use research database composed of aggregated EHR data. Additional steps may be necessary for researchers seeking to implement our methods utilizing data directly collected from the EHR. Grouping related concepts in our dataset was an essential step, otherwise SPADE would not have detected relevant patterns at the specified support levels. Researchers implementing a similar approach may overlook important condition associations if they implement a pattern mining algorithm on EHR data using ICD codes (or similar sparsely used concept codes). Finally, because we used a retrospective observational study design, the associations that SPADE uncovered are purely descriptive. No causation can be attributed to the comorbidities and temporal condition patterns that SPADE identified with pediatric asthma. Rather, this approach should be viewed as a hypothesis-generating method, whose results represent potential associations that future research can investigate for causality.

CONCLUSION

EHR data mining has tremendous potential to improve patient care and reduce financial costs. Frequent pattern mining of EHR data is essential to identify potential associations and correlations in EHR data that researchers may not consider or may have otherwise gone unnoticed. We presented an approach to analyzing a large and complex EHR dataset for temporal condition patterns, using pediatric asthma as a case study. Our analysis revealed strong associations between asthma and several comorbidities and temporal condition patterns. These associations can be used as hypotheses to explore causality in future pediatric asthma research. The methodology presented in this study can be applied to identify temporal patterns in EHR data to investigate conditions and research objectives in numerous contexts outside pediatric asthma.

FUNDING

This work was supported by the Commonwealth Universal Research Enhancement program of the Pennsylvania Department of Health grant number 2015 Formula award SAP #4100072543.

AUTHOR CONTRIBUTIONS

AJM, EAC, and EJB conceived of and designed the study. EAC and AJM conducted the data review. EAC performed the analysis. EAC, AJM, and EJB analyzed the results. EAC wrote the manuscript. All authors contributed to the review and revisions of the manuscript. All authors have seen and approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES:

1. Hand DJ. Principles of data mining. *Drug Saf* 2007; 30 (7): 621–2.
2. Fournier Viger P, Lin C-W, Rage U, *et al*. A survey of sequential pattern mining. *Data Sci Pattern Recognit* 2017; 1: 54–77.
3. Chen F, Deng P, Wan J, *et al*. Data mining for the internet of things: literature review and challenges. *Int J Distrib Sens Netw* 2015; 11 (8): 431047.
4. Batal I. Temporal data mining for healthcare data. In: Reddy CK, Aggarwal CC, eds. *Healthcare Data Analytics*. Boca Raton, FL: CRC Press; 2015: 379–402.
5. Mane RV. A Comparative Study of Spam and PrefixSpan Sequential Pattern Mining Algorithm for Protein Sequences. In: Unnikrishnan S, Serve S,

Bhoir D, eds. *Advances in Computing, Communication, and Control*. Berlin, Germany: Springer; 2013: 147–55.
6. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Bocca JB, Jarke M, Zaniolo C, eds. Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann; 1994: 487–99.
7. Thomas R, Pandy Y. Performance evaluation on state of the art sequential pattern mining algorithms. *Int J Comput Appl* 2013; 65 (14): 8–15.
8. Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. In: Apers P, Bouzeghoub M, Gardarin G, eds. *Advances in Database Technology—EDBT '96*. Berlin, Germany: Springer; 1996: 1–17.
9. Fournier-Viger P, Gomariz A, Campos M, *et al*. Fast vertical mining of sequential patterns using co-occurrence information. In: Tseng VS, Ho TB, Zhou Z-H, Chen ALP, Kao H-Y, eds. *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer; 2014: 40–52.
10. Grover N. Comparative study of various sequential pattern mining algorithms. *Int J Comput Appl* 2014; 90 (17): 36–41.
11. Jian P, Jiawei H, Mortazavi-Asl B, *et al*. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Trans Knowl Data Eng* 2004; 16 (11): 1424–40.
12. Yadav P, Steinbach M, Kumar V, *et al*. Mining electronic health records (EHRs): a survey. *ACM Comput Surv* 2018; 50 (6): 1–40.
13. Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical records with visual analytics. *J Biomed Inform* 2015; 56: 369–78.
14. Gotz D, Wang F, Perer A. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *J Biomed Inform* 2014; 48: 148–59.
15. Chen C, Pai T, Lin S, *et al*. Application of PrefixSpan algorithms for disease pattern analysis. In: International Computer Symposium (ICS); 15–17 December, 2016; Taiwan.
16. Wright AP, Wright AT, McCoy AB, *et al*. The use of sequential pattern mining to predict next prescribed medications. *J Biomed Inform* 2015; 53: 73–80.
17. Wei WQ, Bastarache LA, Carroll RJ, *et al*. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017; 12 (7): e0175508.
18. Sanders R. The Pareto principle: its use and abuse. *J Serv Market* 1987; 1 (2): 37–40.
19. Boytcheva S, Angelova G, Angelov Z, *et al*. Mining comorbidity patterns using retrospective analysis of big collection of outpatient records. *Health Inf Sci Syst* 2017; 5 (1): 3.
20. Bailey LC, Milov DE, Kelleher K, *et al*. Multi-institutional sharing of electronic health record data to assess childhood obesity. *PLoS One* 2013; 8 (6): e66192.
21. Weiner JP, Abrams C. The Johns Hopkins ACG System Technical Reference Guide Version 9.0. 2009. https://www.healthpartners.com/ucm/groups/public/@hp/@public/documents/documents/dev_057914.pdf. Accessed December 4, 2018.
22. Hughes HK, Matsui EC, Tschudy MM, *et al*. Pediatric asthma health disparities: race, hardship, housing, and asthma in a national survey. *Acad Pediatr* 2017; 17 (2): 127–34.
23. Herzog R, Cunningham-Rundles S. Pediatric asthma: natural history, assessment, and treatment. *Mt Sinai J Med* 2011; 78 (5): 645–60.
24. International Classification of Diseases, *Ninth Revision, Clinical Modification (ICD-9-CM)*. Atlanta, GA: Centers for Disease Control and Prevention; 2013. https://www.cdc.gov/nchs/icd/icd9cm.htm. Accessed December 4, 2018]
25. International Classification of Diseases, *Tenth Revision, Clinical Modification (ICD-10-CM)*. Atlanta, GA: Centers for Disease Control and Prevention; 2018. https://www.cdc.gov/nchs/icd/icd10cm.htm. Accessed December 4, 2018.
26. Observational Health Data Sciences and Informatics. *Condition Domain: Observational Health Data Sciences and Informatics*. 2016. http://www.ohdsi.org/web/wiki/doku.php? id=documentation: vocabulary: condition. Accessed December 4, 2018.
27. *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018. https://www.R-project.org/.
28. *Python 3.6.0 Wilmington, DE: Python Software Foundation*; 2016. https://www.python.org/downloads/release/python-360/.
29. Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. *Mach Learn* 2001; 42 (1/2): 31–60.
30. Reshamwala A, Mishra N. Analysis of sequential pattern mining algorithms. *Int J Sci Eng Res* 2014; 5 (2): 1034–8.
31. Kachhadiya BC, Patel B, eds. A survey on sequential pattern mining algorithm for web log pattern data. In: 2nd International Conference on Trends in Electronics and Informatics (ICOEI); 11–12 May 2018; Tamil Nadu, India.
32. Khandagale SMS, Kharat K, Bansode V. Food recommendation system using sequential pattern mining. *Imp J Interdiscip Res* 2016; 2 (6): 912–5.
33. Hahsler MB, Gruen B, Hornik K. arules: Mining Association Rules and Frequent Itemsets. 2018. https://CRAN.R-project.org/package=arules. Accessed April 1, 2018.
34. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014; 2 (1): 3.
35. Healthcare big data and the promise of value-based care. *N Engl J Med Catalyst*. 2018. https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290. Accessed April 6, 2019.
36. Murdoch T, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013; 309 (13): 1351–2.
37. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. *Int J Inf Manage* 2015; 35 (2): 137–44.
38. Healthcare Cost and Utilization Project. *HCUP CCS Fact Sheet*. Rockville, MD: Agency for Healthcare Research and Quality; 2012. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp. Accessed January 13, 2019.
39. Clark NM, Lachance L, Benedict MB, *et al*. The extent and patterns of multiple chronic conditions in low-income children. *Clin Pediatr (Phila)* 2015; 54 (4): 353–8.
40. Matsuno O, Ono E, Takenaka R, *et al*. Asthma and sinusitis: association and implication. *Int Arch Allergy Immunol* 2008; 147 (1): 52–8.
41. Mirabelli MC, Hsu J, Gower WA. Comorbidities of asthma in U.S. children. *Respir Med* 2016; 116: 34–40.
42. Patel MR, Leo HL, Baptist AP, *et al*. Asthma outcomes in children and adolescents with multiple morbidities: Findings from the National Health Interview Survey. *J Allergy Clin Immunol* 2015; 135 (6): 1444–9.
43. de Groot EP, Duiverman EJ, Brand PL. Comorbidities of asthma during childhood: possibly important, yet poorly studied. *Eur Respir J* 2010; 36 (3): 671–8.
44. Jensen AB, Moseley PL, Oprea TI, *et al*. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 2014; 5 (1): 4022.
45. Beck MK, Jensen AB, Nielsen AB, *et al*. Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep* 2016; 6 (1): 36624.
46. Cole AM, Stephens KA, Keppel GA, *et al*. Extracting electronic health record data in a practice-based research network: processes to support translational research across diverse practice organizations. *EGEMS (Wash DC)* 2016; 4 (2): 1206.
47. Kahn MG, Callahan TJ, Barnard J, *et al*. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 1244–62.
48. Batra S, Sachdeva S. Organizing standardized electronic healthcare records data for mining. *Health Policy Technol* 2016; 5 (3): 226–42.