Computer Methods and Programs in Biomedicine 178 (2019) 145–153

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Pediatric population health analysis of southern and central Illinois region: A cross sectional retrospective study using association rule mining and multiple logistic regression

Elham Khorasani Buxton[a,*], Sameer Vohra[b], Yanhui Guo[a], Amanda Fogleman[b], Rushabh Patel[a]

[a] University of Illinois at Springfield, Department of Computer Science, United States
[b] Office of Population Science and Policy, Southern Illinois University School of Medicine, United States

ARTICLE INFO

Article history:
Received 3 September 2018
Revised 21 May 2019
Accepted 17 June 2019

Keywords:
Population health analysis
Data mining
Medical billing database
Pediatric diseases risk factors

ABSTRACT

Background: Southern Illinois University School of Medicine (SIUSOM) collects large amounts of data every day. SIUSOM and other similar healthcare systems are always looking for better ways to use the data to understand and address population level problems. The purpose of this study is to analyze the administrative dataset for pediatric patients served by Southern Illinois University School of Medicine (SIUSOM) to uncover patterns that correlate specific demographic information to diagnoses of pediatric diseases. The study uses a cross-sectional database of medical billing information for all pediatric patients served by SIUSOM between June 2013 and December 2016. The dataset consists of about 980.9K clinical visits for 65.4K unique patients and includes patient demographic identifiers such as their sex, date of birth, race, anonymous zipcode and primary and secondary insurance plan as well as the related pediatric diagnosis codes. The goal is to find unknown correlations in this database.

Method: We proposed a two step methodology to derive unknown correlations in SIUSOM administrative database. First, Class association rule mining was used as a well-established data mining method to generate hypothesis and derive associations of the form $D \rightarrow M$, where D is diagnosis code of a pediatric disease and M is a patient demographic identifier (age,sex, anonymous zipcode, insurance plan, or race). The resulting associations were pruned and filtered using measures such as lift, odds ratio, relative risk, and confidence. The final associations were selected by a pediatric doctor based on their clinical significance. Second,each association rule in the final set was further validated and adjusted odds ratios were obtained using multiple logistic regression.

Results: Several associations were found correlating specific patients' residential zip codes with the diagnosis codes for viral hepatitis carrier, exposure to communicable diseases, screening for mental and developmental disorder in childhood, history allergy to medications, disturbance of emotions specific to childhood, and acute sinusitis. In addition, the results show that African American patients are more likely to be screened for mental and developmental disorders compared to White patients for SIUSOM pediatric population (Odds Ratio (OR):3.56, 95% Confidence Interval (CI):[3.29,3.85]).

Conclusion: Class association rule mining is an effective method for detecting signals in a large patient administrative database and generating hypotheses which correlate patients' demographics with diagnosis of pediatric diseases. A post processing of the hypotheses generated by this method is necessary to prune spurious associations and select a set of clinically relevant hypotheses.

1. Introduction

Southern Illinois University School of Medicine's (SIUSOM) mission is to assist the people of central and southern Illinois in meeting their health care needs through education, patient care, research, and service to the community. The last priority is relatively

* Corresponding author.
E-mail addresses: esahe2@uis.edu (E.K. Buxton), svohra97@siumed.edu (S. Vohra), yguo56@uis.edu (Y. Guo), afogleman@siumed.edu (A. Fogleman), rpate307@uis.edu (R. Patel).

https://doi.org/10.1016/j.cmpb.2019.06.020
0169-2607/

new addition to SIUSOM. Service to the community expands the medical school's goals beyond the traditional confines of the clinics and medical schools and requires its physicians, medical students, and researchers to improve health by understanding the populations it serves. SIUSOM serves 66 counties in central and southern Illinois. This service region constitutes over 32,000 square miles and over two million people. To better serve the community, SIUSOM must better understand its region, and this starts with better understanding the data and information of the patients it serves. However, SIUSOM has never tried to view the data in sum to understand if there are any unique patterns uncovered that correlates specific demographic information to disease and disease outcome. This type of data analytics could help SIUSOM pinpoint the prevalence of certain disease to specific communities and take proactive measures to better serve the community.

For this study, we used the medical billing data for all pediatric population served by SIUSOM between June 2013 and December 2016. The goal was to examine the relationship between patient demographic identifers (age, sex, zipcode, race, primary insurance plan) and diagnosis of pediatric diseases in the pediatric population served by SIUSOM.

We applied a two-step method to discover and test unknown patterns in the SIUSOM administrative dataset with no prior hypothesis. In the first step, Class Association Rule Mining (CARM) [1] was used to generate hypothesis and detect patterns of the form $D \rightarrow M$ where M is a demographic factor and D is a diagnosis group for a pediatric disease based on International Classification of Diseases ICD-9 and ICD-10 (note that "$\rightarrow$" should not be interpreted as a causal relationship). The resulting association rules were evaluated and filtered using standard disproportionality measures [2] such as Odds Ratio (OR), Relative Risk (RR), and confidence. Final associations were selected by a medical expert based on their clinical significance. In the second step, each association rule was further examined by looking at the specific codes in the diagnosis group and the sub-population to which it is applied. Then a multiple logistic regression was used for each association to estimate the effect of the demographic factor in the diagnosis of the pediatric disease and to adjust for confounders.

The rest of the paper is organized as follows: Section 2 briefly reviews the related work. Section 3 describes hypothesis generation from administrative data using CARM and discusses association rule evaluation and pruning. Section 4 validates and quantifies the strength of the final set of clinically interesting association rules using multiple logistic regression. The last section gives concluding remarks and discusses the strengths and the limitations of the study and its future research direction.

2. Related work

CARM have been used on Electronic Health Records as well as administrative billing data to identify associations among clinical concepts and to generate hypothesis [3]. Kang'ethe and Wagacha [4] used Association Rule Mining on a dataset of 98,000 patient clinic visits to extract diagnosis codes which co-occur frequently. Wright et. al. [5] performed Association Rule Mining on structured electronic health records of a sample of 100,000 patients in order to identify associations between medications, laboratory results and problems. Concaro et. al. [6] used temporal Association Rule Mining on administrative and clinical data to find associations between clinical variables and drug effects. There have been several studies on using association rule mining in pharmacovigilance for detecting adverse drug effects in spontaneous adverse event reporting system [7–9] Association Rule mining has also been used in epidemiology to identify risk factors for several diseases including early childhood caries [10], hypertension [11], type 2 diabetes [12], heart disease [13], cancer [14], and general lifestyle risk behaviours [15].

While CARM has been previously used to find risk factors for specific diseases, to our knowledge, it has never been used to identify risk factors across all diseases in an administrative billing database.

One of the shortcomings of CARM is that it could generate a large number of spurious associations. Therefore, often post-processing of the results are needed to filter out true and clinically interesting associations. Studies in medical application of CARM, typically use standard measures such as confidence, lift, chi-square or disproportionality measures such as odds ratio or relative risk to select relevant rules. While these measures are helpful to prune the rules, they do not filter out confounded associations. For example, the association between a disease and patients' location might be confounded by the race of the patients. Multiple logistic regression is one of the most widely used methods in epidemiology to measure the effect of one or more risk factors on development or diagnosis of a disease while adjusting for confounding variables [16,17].

The contribution of this study is using CARM to generate hypotheses which relate patient demographics to diagnosis of various pediatric diseases in a large medical billing database. We do not stop at detecting the meaningful associations but further examine each association rule by breaking down its diagnosis group and apply multiple logistic regression to the relevant population to account for confounding and to obtain adjusted odd ratios.

3. Methodology

3.1. Background

Association Rule Mining is a common data mining method for discovering hidden relationships between variables in a large database. Let $I = \{I_1, I_2, \ldots\}$ be a set of items in the database and $\{D \subset I\}$ is a set of transactions. An association rule is an implication of the form $A \rightarrow B$ where A, $B \in I$ and $A \cap B = \emptyset$. The strength of an association rule is typically measured in terms of its *support* and *confidence*.

The support of a rule $A \rightarrow B$ is the percentage of transactions in D which contain both A and B, that is, how often A and B co-occur in the database.

$$support(A \rightarrow B) = \frac{|\{d \in D; A, B \subseteq d\}|}{|D|}$$

The confidence of a rule $A \rightarrow B$ is an estimate of the conditional probability $P(B|A)$ and measures the percentage of transactions in D with A in them which also contain B.

$$confidence(A \rightarrow B) = \frac{support(A \rightarrow B)}{support(B)}$$

Where $suppor(B)$ is the ratio of all transactions in the database which contain B.

The common association rule mining algorithm, Apriori [18], discovers association rules in a large database in two main steps:

1. *Frequent itemset generation*: it finds all combinations of items that have transaction support above a minimum, user-defined, threshold. To do so, Apriori takes advantage of the anti-monotonicity property of the support measure, that is, all subsets of a frequent itemset are also frequent and all superset of an infrequent itemset are also infrequent. Apriori starts with finding all frequent itemsets with a single item. It proceeds by constructing frequent itemsets with $k+1$ items by self-joining frequent itemsets with k items and continues the same process until no new frequent itemset can be generated.

Table 1
Description of variables used to derive associations. The percentages in the last column show the relative frequency for each level.

Variable	Description	Levels & distribution
Age	Age group of the patient at the time of visit	Infant:23.4%, Toddler:6.5%, Early childhood:13%, Middle childhood:22%, Early adolescence:35.1%
Gender	Sex of the patient	Fenale:48.4%, Male:51.6%,
Race	Race of the patient	White:77.1%, Black/African American:16%, Mixed Race:1.6%, Asian:0.7%, Other:4.6%
Primary Insurance	Patient's primary insurance plan	Medicaid:54.5%, Managed Care:43%, Self pay/No insurance:1.5%, Commercial:0.8%, Other:0.2%
Geo-code	Patient's de-identified zipcode	876 levels: 1381:8.6%, 1382: 8.2%, 1383:7%, 1302:3.5%, etc.
ICD	Diagnosis group for the clinical visit	1184 levels: Z23(Encounter for immunization):4.5%, Z38:(liveborn infant):2.7%, R06:(abnormalities of breathing):2.7%, etc.

2. *Rule Generation* Once frequent itemsets are found, Apriori partitions a frequent itemset to form the left-hand-side and right-hand-side of a rule and keeps the rules which satisfy a user-defined threshold for confidence.

3.2. Data preparation

The data source used in this study consist of patient information (de-identified patient id, date of birth, anonymous zip code,race, primary insurance category and name, secondary insurance category and name, provider name, service date, and diagnosis code and description) for all pediatric visits to SIUSOM facilities between June 2013 and December 2016. The study was approved by SIU Institutional Review Board.

Before applying Aprior algorithm to discover patterns in this dataset, several preprocessing steps were necessary to prepare the data and ensure its consistency. These steps are summarized below:

1. Patients' age at the time of diagnosis was calculated from date of birth and service date. Age was then discretized to four categories based on the age groups recommended in [19]: infant (0–12 months)year, toddler(13 months-2 years), early childhood (2–5 years), middle childhood (6–11 years), and early adolescence(12–18 years).
2. Patients' race was categorized to "White", "African-American", "Asian", "Mixed Race", "Unknown", and "Other".
3. The dataset contains both ICD9 and ICD10 diagnosis codes. To draw valid associations, the diagnosis codes must be made consistent across the dataset. For example, asthma should either be coded as 493.XX (ICD 9) or J45.XX (ICD10) but not both.'We converted all ICD9 codes to ICD10 using the CMS General Equivalence Mapping (GEM) database [20] and the R package icdcoder [21].
4. The diagnosis group was extracted from each ICD code. This is because, the characters after period (.) in ICD9 and ICD10 represent the etiology, anotomic site, or severity of the disease category. For example 493, and J45 are the general ICD9 and ICD10 codes for asthma, respectively while ICD9 code 493.10 ("Mild intermittent asthma, uncomplicated") and ICD10 code J45.50 ("Severe persistent asthma, uncomplicated") are more detailed codes for asthma representing its severity. This generalization is necessary to avoid treating multiple specific codes for the same disease as separate diseases and to derive associations on a higher level.
5. Redundant records with the same patient id and diagnosis code were removed from the database. If a patient had multiple visits for the same diagnosis code, only the earliest visit was retained and all his/her subsequent visits for the same diagnosis code were removed. This step is necessary to avoid over-representing the same transaction in the database and to ensure that the patient age is correctly computed at the first diagnosis of the disease.A major portion of data (about two-third) was redundant and eliminated in this step.
6. Each visit was transformed into (age, gender, race, geo-code, primary insurance category, diagnosis group) to form a transaction. Geo-codes are patients' masked zipcodes substituted by random values for HIPPA compliance. All transactions with a missing value for any of these attributes where removed from the databases (The deleted records constituted about 1% of the database). The description of each variable and its levels are listed in Table 1.

After preprocessing, the database was reduced to 302,882 clinical visits for 64,241 unique pediatric patients served by SIUSOM between June 2013 and December 2016.

3.3. Association rule/hypothesis generation and pruning

One could theoretically run a multiple logistic regression for each diagnosis code to find the effect of patients' demographics variables on the particular disease represented by this code. However, such approach would be time consuming and not scalable. An alternative, more efficient approach is to use class association rule mining to generate an initial set of hypotheses and possible associations and then examine each interesting association using multiple logistic regression.

Class association rules are a subset of association rules in which the right hand side of the rule belongs to a specified class. In the context of this study Class Association Rule Mining is the problem of finding association rules of the form $D \rightarrow M$ where M is a patient's demographic identifier (race, age, gender, anonymous zip code, and primary insurance category) and D is a diagnosis group for a pediatric disease (the arrow $\rightarrow$ should not be interpreted as causal relationship).

We used R package *arules* [22] to generate association rules with support values greater than or equal to 50 transaction (i.e., at least 50 patients with each diagnosis group) and confidence of 1%. There is no universal threshold for support and confidence and they vary by the study domain. A high support threshold may cause associations for infrequent rare diseases to go undetected. A high confidence threshold might only choose rules which might be statistically significant but clinically trivial and ignore more clinically interesting rules with lower confidence [23]. For example, the rule $ICD = P92* \rightarrow Age = Infant$ (where ICD P92*=feeding problem of newborn) is clinically trivial but has a maximum confidence value of 1.0 because P92* is only billed for infants. In addition it is well-known that a high confidence does not always mean a strong association as a rule with a frequent item on the right-hand-side is likely to have a high confidence regardless of the item on the left-hand side and a rule with a rare item on the right hand side is likely to generate a low confidence regardless of its association with the left-hand side. For these reasons, we experimented with a range of values for support and confidence thresholds and heuristically chose the values which generated the maximum set of valid rules with a lift value of at least 2.

Lift of a rule $D \rightarrow M$ is the ratio of the joint probability $P(M \cap D)$ over the expected probability of P and M occurring together under

Table 2
Patient demographic-disease contingency table.

	M	¬M
D	$\|M \cap D\|$	$\|\neg M \cap D\|$
¬D	$\|M \cap \neg D\|$	$\|\neg M \cap \neg D\|$

the independence assumption. Lift addresses the bias towards frequent right hand side by dividing the confidence of the rule by the probability of the right hand side:

$$lift(D \rightarrow M) = \frac{P(M|D)}{P(M)} = \frac{P(M \cap D)}{P(M) \times P(D)}$$

Large values of lift indicate that M and D are positively correlated and their co-occurrence is unlikely due to chance while lift=1 indicates that there is no correlation between M and D. Similar to Hanauer and co-workers [7,24,25] we used a lift threshold equal to 2 to prune spurious rules. We examined the remaining rules and pruned the associations with a confidence greater than or equal to 50%. Although statistically significant, these associations were not clinically interesting. For instance, the trivial association rule $icd = P95^*(neonatal\,jaundice) \rightarrow Age = Infant)$ has a confidence value equal to 1. Similarly, associations between sickle cell, dermatitis, and umbilical hernias diseases and race=Black/African American all have high confidence values (greater than 55%) but represent a common knowledge in medicine and hence are not clinically interesting.

Additional measures such as *odds ratio* and *relative risk* were used to further evaluate, rank, and quantify the interestingness of the association. Odds ratio and relative risk are two common metrics used in epidemiology to measure the strength of association between exposures and outcomes [26]. Odds ratio is the odds of disease in the exposure group divided by the odds of disease in non-exposure group. Relative risk is the probability of disease in the exposure group divided by the probability of disease in the non-exposed group.

To calculate odds ratio and relative risk of an association rule $D \rightarrow M$, we need to compute the contingency table shown in Table 2 for each patient demographic-disease association: where $|M \cap D|$ is the number of patients in the data set with demographic M who were visited for disease D, $|\neg M \cap D|$ is the number of patients with demographic other than M who were visited for disease D. $|M \cap \neg D|$ is the number of patients with demographic M who were not visited for disease D, and $|\neg M \cap \neg D|$ is the number of patients with demographic other than M who were visited for any disease other than D during the observation period.

The odds ratio of the rule $D \rightarrow M$ is the ratio of the odds of a patient visited for disease D having demographic M to the odds of a patient visited for D not having demographic M.

$$OR = \frac{odd(D|M)}{odd(D|\neg M)} = \frac{\frac{p(D|M)}{p(\neg D|M)}}{\frac{p(D|\neg M)}{p(\neg D|\neg M)}} = \frac{|D \cap M| \times |\neg D \cap \neg M|}{|\neg D \cap M| \times |D \cap \neg M|}$$

The distribution of the log odds ratio is approximately normal with the standard error approximately equal to:

$$SE(log(OR)) = \sqrt{\frac{1}{|D \cap M} + \frac{1}{D \cap \neg M} + \frac{1}{\neg D \cap M} + \frac{1}{\neg D \cap \neg M}}$$

and hence, the 95% confidence interval for odds ratio is computed as:

$$CI(OR) = exp(log(OR) \pm 1.96 \times SE(log(OR)))$$

If the lower value of CI(OR) is greater than one, it suggests that D and M are positively correlated.

The relative risk of the rule $D \rightarrow M$ is the ratio of the probability of a patient with demographic M visited for disease D to the probability of a patient with demographics other than M visited for disease D.

$$RR = \frac{P(D|M)}{P(D|\neg M)} = \frac{\frac{p(D \cap M)}{p(M)}}{\frac{p(D \cap \neg M)}{p(\neg M)}} = \frac{\frac{|D \cap M|}{|D \cap M| + |\neg D \cap M|}}{\frac{|D \cap \neg M|}{|D \cap \neg M| + |\neg D \cap \neg M|}}$$

The distribution of the log relative risk is approximately normal with standard error approximately equal to:

$$SE(log(RR)) = \sqrt{\frac{1}{|D \cap M|} + \frac{1}{|D \cap \neg M|} - \frac{1}{|D \cap M| + |\neg D \cap M|} - \frac{1}{|D \cap \neg M| + |\neg D \cap \neg M|}}$$

and hence, the 95% confidence interval for relative risk is computed as:

$$CI(RR) = exp(log(RR) \pm 1.96 \times SE(log(RR)))$$

If the lower value of CI(RR) is greater than one it indicates that there is an increased risk of having a visit for disease D for patients with demographic M.

After initial screening and pruning of several clinically trivial associations, a set of 14 association rules were identified by a medical expert as being clinically interesting.[1] These association rules, together with their lift, confidence and lower 95% confidence bounds for odds ratio and relative risk are shown in Table 3. As shown in the table, many of the clinically significant rules are the ones that correlate patients' geo-codes (i.e., masked patients' zip codes) to diagnosis of a particular disease. The counties where these geo-codes are located are also shown in the table. There are a couple of associations indicating African Americans are more likely to be screened for V79* (mental and development disorder in childhood) and uninsured patients have a slightly increased risk for Neonatal Jaundice. These associations merely signal possible relationship and should be further examined and validated as explained in the next section.

4. Testing associations

While CARM can efficiently detect unknown correlations between diseases and patients' demographics in SIUSOM pediatric database, these correlations do not necessarily mean definite relationships and should be further examined. First, the specific ICD9 and ICD10 codes in the diagnosis group of each association rule should be explored to find out whether there are any specific codes which account for most of the correlation in each association rule. Second, some diagnosis codes might only be applicable to a specific patient group. For instance, ICD code Z34* (encounter for supervision of routine pregnancy) is typically only recorded for female patients and hence all male patients should be excluded from the analysis of this diagnosis code. Third, CARM does not account for confounding effects. For instance the effect of patients' race could be confounded by their socio-economic status (SES). We use multiple logistic regression to adjust for such confounding effects and to obtain adjusted odds ratios. SIUSOM administrative database does not contain information on patients' education, income, or their occupation; therefore, patient zipcode and their type of insurance (public vs private insurance, or uninsured) as proxies for patient socio-economic status. In what follows, we

[1] CARM detected over 100 association rules with support>= 50 patients, confidence>= 1% and lift>= 2. Many of these rules were trivial rules and were pruned after initial screening. A set of 40 rules were then presented to the medical expert who identified 14 of then as clinically significant. Although the rest of the rules were relevant they were general associations that were already known in the medical community.

Table 3
Association rules generated by CARM after post-processing and pruning.

Rule	Association rule	Diagnosis group description	Lower bound odds ratio	Lower bound relative risk	Lift	Confidence
1	icd=V02* → Masked Zipcode=1302 (Macon County, IL)	Carrier or suspected carrier of infectious' diseases	12.2	11.6	9.6	0.34
2	icd=V02* → Masked Zipcode=1299 (Macon County, IL)	Carrier or suspected carrier of infectious diseases	9.3	8.9	8.2	0.27
3	icd=V02* → Geo-code=1300 (Macon County, IL)	Carrier or suspected carrier of infectious diseases	8.4	8.1	9.2	0.14
4	icd=V01* → Geo-code=1299 (Macon County, IL)	Contact with or exposure to communicable diseases	10.5	9.7	8.5	0.28
5	icd=V01* → Geo-code=1302 (Macon County, IL)	Contact with or exposure to communicable diseases	9.9	9.2	8	0.28
6	icd=V01* → Geo-code=1300 (Macon County, IL)	Contact with or exposure to communicable diseases	7.3	6.9	7.7	0.12
7	icd=V79* → Race=AFRICAN AMERICAN ref= "White"	Special screening for mental disorders and developmental handicaps	3.9	3.4	2.6	0.41
8	icd=V79* → Geo-code=1381 (Sangamon County, IL)	Special screening for mental disorders and developmental handicaps	4.6	3.9	3.3	0.28
9	icd=V79* → Geo-code=1382 (Sangamon County, IL)	Special screening for mental disorders and developmental handicaps	4.3	3.7	3.2	0.26
10	icd=313* → Geo-code=1236 (Adams County, IL)	Disturbance of emotions specific to childhood and adolescence	4.4	4.3	4.7	0.16
11	icd=V28* → Race=AFRICAN AMERICAN	Encounter for antenatal screening of mother	3.2	3.2	2.9	0.46
12	icd=774* → PrimaryInsCategory=SELF PAY	Neonatal jaundice from other and unspecified causes	3.1	3.0	3.7	0.06
13	icd=V14* → Geo-code=1387 (Sangamon County, IL)	Personal history of allergy to medicinal agents	2.6	2.5	3.3	0.05
14	icd=461* → Geo-code=1343 (Sangamon County, IL)	Acute sinusitis	2.6	2.5	3.1	0.04

further examine each association rule in Table 3 by breaking down the diagnosis group to its specific diagnosis codes and apply multiple logistic regression to obtain adjusted odd ratios.

4.1. Testing associations rules 1–6

Association rules 1–6 in Table 3 correlate specific patients' geo-codes with diagnosis groups V02* and V01*.

Diagnosis group V02* indicates patients who are carriers or suspected carriers of infectious disease. The break down of this code together with the number of unique patients diagnosed with them for geo-codes 1302, 1299, and 1300 are shown in Table 4. As can be seen from this table, viral hepatities carrier (diagnosis codes V02.60 and Z22.50) constitutes over 99% of the diagnosis group V02 for all geo-codes. Hence, it is reasonable and more informative to test the association rules 1–3 in Table 3 only for viral hepatitis.

Multiple logistic regression is used to test the associations between viral hepatitis and each of the geo-codes 1302, 1299 and 1300. For instance, to test the effect of the geo-code 1302 on being a viral hepatitis carrier, we use a binary dependent variable indicating whether the patient had the status code V02.60 or Z22.50 during the observation period. Similarly, a binary variable is used as the independent variable which takes the value of one if the patient's geo-code is equal to 1302 and zero otherwise. We also adjust for patients' race and their primary insurance plan. Table 5 reports the 95% confidence intervals and p-values for the estimates of the effect of geo-code 1302, 1299, and 1300 on viral hepatitis carrier status. This table shows that on average the odds of being a viral hepatitis carrier was increased by a factor of 13.9, 10.45 and 9.09 for SIU pediatric patients with geo-codes 1302, 1299 and 1300, respectively.

Diagnosis group V01* shows patients who had exposure to a communicable disease. Table 6 shows the break down of V01* together with the number of unique patients diagnosed with them for geo-codes 12,991,302 and 1300 (rules 4–6 in Table 3).

Multiple logistic regression was used to test the associations between the general status group V01* and each of the geo-codes 1302, 1299 and 1300. The dependent variable takes a binary value indicating whether the patient had any of the status codes in Table 6 and the independent variable for each geo-code is a binary variable indicating whether the patient resides in that geo-code. As before, we adjust for race and socio-economy status proxied by type of the patients' primary insurance plan. Table 7 reports the 95% confidence intervals and p-values for the estimates of the effect of geo-codes 1302, 1299, and 1300 on the status group V01*. This table shows that on average SIU pediatric patient with geo-codes 1299, 1302 and 1300 had an increased odds of exposure to communicable diseases by a factor of 8.54, 7.71 and 6.16, respectively.

4.2. Testing associations rules 7–9

Rules 7–9 in Table 3 correlate the diagnosis group V79*(Screening for mental and developmental disorder) with being an African American as well as the geo-codes 1381 and 1382. Table 8 shows the breakdown of the diagnosis group V79* in the dataset for African American patients as well as patients with geo-codes 1381 and 1382.

The diagnosis code V79.0 (Screening for Depression) is mostly used in the dataset for screening for maternal depression and is irrelevant to the rest of the codes in the diagnosis group V79*. Hence, it is excluded when testing the association rules 7–9. To test each association, a multiple logistic regression with a binary dependent variable is used indicating whether the patient had any of the diagnosis codes (V79.8, Z13.4, V79.3, or V79.9). To test the effects of each of the geo-codes 1381 and 1382 in the diagnosis

Table 4
Break down of the diagnosis group v02 and their patient frequencies for geo-codes 1302, 1299, and 1300.

breakdown of V02 for geo-code 1302		
Diagnosis Code	Diagnosis Description	Patient Frequency
V02.60 or Z22.50	Carrier of unspecified viral hepatitis	147
V02.0	CARRIER CHOLERA	1
Z22.39	Carrier of other specified bacterial diseases-Z22.39	1
breakdown of V02 for geo-code 1299		
Diagnosis Code	Diagnosis Description	Patient Frequency
V02.60	UNSPEC VIRAL HEPATITIS CARRIER	118
V02.54	CARRIER OF MRSA	3
Z22.50	Carrier of unspecified viral hepatitis-Z22.50	1
breakdown of V02 for geo-coden 1300		
Diagnosis Code	Diagnosis Description	Patient Frequency
V02.60	UNSPEC VIRAL HEPATITIS CARRIER	60
V02.0	CARRIER CHOLERA	1

Table 5
The effect of geo-codes 1302, 1299, and 1300 on viral hepatitis carrier status.

Gelocation code	Odds ratio	95% Confidence interval for odds ratio	textbfP-value
1302	13.94	[11.24, 17.24]	$< 2e-16$
1299	10.45	[8.32, 13.05]	$< 2e-16$
1300	9.09	[6.74, 12.07]	$< 2e-16$

Table 6
break down of the diagnosis group v01* and their patient frequencies for geo-codes 1299, 1302 and 1300.

breakdown of V01* for geo-code 1299		
Diagnosis Code	Diagnosis Description	Patient Frequency
V01.79	Contact/exposure to viral diseases	109
V01.84 or Z20.811	Contact/Exposure to meningococcus	79
Z20.6	Contact/exposure to human immunodef virus	1
Z20.89 or V01.89	Contact/exposure to other communicable diseases	2
Z20.5	Contact/exposure to viral hepatitis	1
breakdown of V01* for geo-code 1302		
Diagnosis Code	Diagnosis Description	Patient Frequency
V01.79	Contact/exposure to viral diseases	102
V01.84 or Z20.811	Contact/Exposure to meningococcus	81
Z20.5	Contact with and (suspected) exposure to viral hepatitis-Z20.5	3
Z20.6	Contact/exposure to human immunodef virus	1
Z20.818	Contact/exposure to other bacterial communicable diseases	1
Z20.1	Contact/exposure to tuberculosis	1
breakdown of V01* for geo-code 1300		
Diagnosis Code	Diagnosis Description	Patient Frequency
V01.79	Contact/exposure to viral diseases	38
V01.84 or Z20.811	Contact/Exposure to meningococcus	38
Z20.2	Contact/exposure to infect w a sexl mode of transmiss-Z20.2	2
Z20.6	Contact/exposure to human immunodef virus	1
Z20.828	Contact/exposure to other viral communicable diseases	1
Z20.89 or V01.89	Contact/exposure to other communicable diseases-Z20.89	1

Table 7
The effect of geo-codes 1302, 1299, and 1300 on status group v01*.

Gelocation code	Odds ratio	95% Confidence interval for odds ratio	P-value
1299	8.54	[7.17, 10.14]	$< 2e-16$
1302	7.71	[6.46, 9.15]	$< 2e-16$
1300	6.16	[4.77,7.84]	$< 2e-16$

group V79*, an independent variable was used for each geo-code indicating whether a patient has that geo-code adjusted by patients' race and their primary insurance category. To test the effect of race, a multiple logistic regression was used with race as independent variable (race=WHITE as reference category), adjusted by socioeconomic status proxied by the primary insurance category.

Table 9 reports the 95% confidence intervals and p-values for the estimates of the effect of geo-codes 1381, 1382, and Race=African American on the status group V79*. This table shows that on average the odds of screening for mental or developmental disorder for SIU pediatric patient population increased by a factor of 3.65 and 2.41 or patients with geo-codes 1381 and 1382, respectively. Also the table shows that on average the odds of screening for mental and dvelopmental disorder for SIUSOM African American patients were 3.56 times higher than that of the White patients.

4.3. Testing associations rules 10–14

The diagnosis group 313* indicates behavioural and emotional disorders specific to childhood and adolescence. The breakdown of this diagnosis group for patients with geo-code 1236 is shown

Table 8
Break down of the diagnosis group v79* and their patient frequencies for African American patients and patients with geo-codes 1381 and 1382.

breakdown of V79* for African American Patients		
Diagnosis code	Diagnosis description	Patient frequency
V79.8	Screening For Mental Disorder Other	1208
Z13.4 or V79.3	Encntr screening for certain developmental disorders in childhood	362
V79.0	Screening For Depression	266
V79.9	SCREENING FOR Mental Disorder Unspecified	3
breakdown of V79* for geo-code 1381		
Diagnosis Code	Diagnosis Description	Patient Frequency
V79.8	Screening For Mental Disorder Other	825
Z13.4 or V79.3	Encntr screen for certain developmental disorders in childhood	259
V79.0	Screening For Depression	189
V79.9	SCREENING FOR Mental Disorder Unspecified	3
breakdown of V79* for geo-code 1382		
Diagnosis Code	Diagnosis Description	Patient Frequency
V79.8	Screening For Mental Disorder Other	756
Z13.4 or V79.3	Encntr screen for certain developmental disorders in childhood	215
V79.0	Screening For Depression	181

Table 9
The effect of geo-codes 1381, 1382, and Race=African American on status group v79*.

Independent variable	Odds ratio	95% Confidence interval for odds ratio	P-value
Geo Code==1381	3.65	[3.35, 3.98]	$< 2e-16$
Race=African American Ref="White"	3.56	[3.29, 3.85]	$< 2e-16$
Geo Code==1382	2.41	[2.19, 2.65]	$< 2e-16$

in Table 10. Considering that over 85% of the cases are ICD codes 313.81 or F91.3 (oppositional defiant disorder), it makes sense to only test the association for these diagnosis codes instead of the broader category 313*. To estimate the effect of geo-code 1236 on the diagnosis of oppositional defiant disorder, we run a multiple logistic regression with a binary dependent variable indicating whether a patient had the icd group 313* during the observation period and a binary independent variable indicating whether the patient had geo-code 1236 while adjusting for race and patients' socioeconomic status proxied by their insurance plan. The resulting estimates are shown in Table 11. On average, the odds of diagnosis with oppositional defiant disorder for SIUSOM pediatric patients increased by a factor of 5.69 for patients with geo-code 1236.

Diagnosis group V28* indicates prenatal screening for pregnant women. In the context of pediatric population, it shows adolescent pregnancy. The breakdown of this diagnosis group for African American adolescent patients in SIU dataset is shown in Table 10. The association between race=African American and diagnosis group v28* in SIU pediatric dataset signals higher teen pregnancy rates for African American patients. This is in accordance with data from US Department of Health and Human Services, office of Adolescent Health [27]. The effect of race (adjusted by socioeconomic status proxied by insurance plan) on the diagnosis group V28* for SIUSOM adolescent patients is shown in Table 11, on average, the odds of teen pregnancy for African American patients is 3.47 times higher than White patients for SIUSOM adolescent pediatric population.

Diagnosis group 774* indicates neonatal jaundice. Rule 12 in Table 3 indicates a possible association between being uninsured and the diagnosis of neonatal jaundice among SIUSOM pediatric patients. The breakdown of the diagnosis group 774* for uninsured patients is shown in Table 10. To test the effect of being uninsured on neonatal jaundice in SIU pediatric population, we use a logistic regression with a binary dependent variable, indicating whether a newborn patient has been diagnosed with nenonatal jaundice during the observation period. The independent variable is also binary indicating whether a newborn was uninsured. As shown in Table 11 the effect of being uninsured on newborn neonatal jaundice for SIUSOM newborn patients is not very significant[2] (confidence interval [1.02,1.73] and p-value 0.03). This result is inconsistent with Table 3 where a higher odds ratio and relative risk were reported for this association. This is because the association rule mining merely generates hypothesis based on the disproportional co-ocurrences of a diagnosis code and demographic identifier disregarding the context in which the diagnosis code is applied. In the case of diagnosis group 774*, CARM compares the newborns with 774* to all other patients from all age groups who do not have this diagnosis code. This can lead into generating false hypothesis about a wrong population as neonatal jaundice is only applicable to the newborn age group. Hence, CARM should only be used as a technique that detects signals in a large database. The hypothesis generated from CARM should further be tested based on the context of the diagnosis group and the population to which it is applied.

Rule 13 in Table 3 shows a possible correlation between patient geo-code 1387 and status group V14* indicating personal history of allergy to medication. The breakdown of status group V14* and their patient frequency for geo-code 1387 is shown in Table 10. To test this correlation, a multiple logistic regression was used with a binary dependent variable indicating whether the patient was billed for the status group V14* (or Z88* in ICD 10). The independent variable is a binary variable indicating whether a patient has geo-code 1387. The effect of geo-code 1387 on status group V14 adjusted by patients' race and their insurance plan is presented in Table 11. On average, the odds of the status group V14*(Personal history of allergy to medicinal agents) is increased by a factor of 2.07 for SIUSOM pediatric patients with geo-code 1387.

Rule 14 in Table 3 shows a possible correlation between patient geo-code 1343 and diagnosis group 461*(or J01*, ICD 10) acute sinusitis. The breakdown of status group 461* and their patient frequency for geo-code 1323 is shown in Table 10. To test this correlation, a multiple logistic regression was used with a binary dependent variable indicating whether the patient was billed for

[2] at the significance level=0.01.

Table 10
Break down of the diagnosis groups 313*,V28*, and 774* and their patient frequencies for patients with geo-code 1236, African American patients and uninsured patients, respectively.

breakdown of 313* for patients with geo-code=1236		
Diagnosis code	Diagnosis description	Patient frequency
313.81 or F91.3	Oppositional defiant disorder	127
F91.9	Conduct disorder, unspecified	7
F94.1	Reactive attachment disorder of childhood	5
313.89	OTH EMOTIONAL DISTURBANCE CHILDHOOD	4
F98.9	Unspecified childhood behavioural and emotional disorder	2
F98.8	Other Specified childhood behavioural and emotional disorder	2
F91.8	Other conduct disorders	1
breakdown of V28* for Black/African American Patients		
Diagnosis Code	Diagnosis Description	Patient Frequency
V28.3	Screening for fetal malformation	54
Z36	Encounter for antenatal screening of mother	11
V28.9	Antenatal Screening For unspecified	10
V28.6	Antenatal Screen Srep B	5
V28.89	Other Antenatal Screening	1
V28.0	Screening for chromosome anomaly	1
V28.5	Screening for isoimmunization	1
breakdown of 774* for uninsured Patients		
Diagnosis Code	Diagnosis Description	Patient Frequency
P59.9 or 774.6	Neonatal jaundice, unspecified	64
774.2 or P59.0	Neonatal jaundice pre-term delivery	7
P59.3	Neonatal jaundice from breast milk inhibitor	1
breakdown of V14* for patients with geo-code 1387		
Diagnosis Code	Diagnosis Description	Patient Frequency
Z88.0 or V14.0	Allergy status to penicillin-Z88.0	74
Z88.1 or V14.1	Allergy status to other antibiotic agents	39
Z88.8 or V14.8	Allergy status to other drugs/meds/biol substances	19
V14.2	Personal history of allergy to sulfonamides	13
Z88.5,Z88.6,Z88.4,V14.4,V14.5,or V14.7	Allergy status to analgesic,anesthetict, narcotic, or serum vaccine	13
breakdown of 461* for patients with geo-code 1343		
Diagnosis Code	Diagnosis Description	Patient Frequency
461.9 or J01.90	Acute sinusitis unspecified	79
J01.00 or 461.0	Acute maxillary sinusitis	6
Z01.80	Other acute sinusitis	1

Table 11
The effect of geo-code 1236 on diagnosis of oppositional defiant disorder, the effect of race=black/African American on diagnosis group V28*, and the effect of insurance category=uninsured on diagnosis group P59*.

The effect of geo-code 1236 on diagnosis of oppositional defiant disorder			
Independent variable	Odds ratio	95% Confidence interval for odds ratio	P-value
Geo Code==1236	5.69	[4.52,7.09]	$< 2e-16$
The effect of Race=African American on the disagnosis group V28*			
Race=African American (ref="White")	3.47	[2.44,4.94]	$6.02e-12$
The effect of insurance category=uninsured on the diagnosis group 774*			
Primary Insurance==Self Pay/Uninsured	1.34	[1.02,1.73]	0.03
The effect of geo-code 1387 on the status group V14*			
Geo code==1387	2.07	[1.64,2.58]	$2.3e-10$
The effect of geo-code 1343 on the status group 461*			
Geo code==1343	2.54	[2.13,3.01]	$< 2e-16$

the diagnosis group 461*(or J01*, ICD 10). The independent variable is a binary variable indicating whether a patient has geo-code 1343. The effect of geo-code 1343 on acute sinusitis adjusted by patients' race and their socioeconomic status proxied by their insurance plan is presented in Table 11. On average, the odds of the diagnosis with acute sinusitis for SIU pediatric patients is increased by a factor of 2.54 for patients with geo-code 1343.

5. Discussion and future work

Analysis of administrative data helps SIUSOM better understand the risk factors for pediatric patients it serves in Central and Southern Illinois. The contribution of the study presented here is a two step analysis method for finding unknown correlations between patient demographics and diagnosis of pediatric diseases in a large SIUSOM administrative database. First, CARM was used for detecting signals and generating hypotheses and then multiple logistic regression was used to compute adjusted odds ratios for each hypothesis.

Several associations were found in this study correlating specific patients' zipcodes with diagnosis of some pediatric diseases. It would be desirable to understand the characteristics of the communities with those zipcodes and their effect on the diagnosis of pediatric diseases. To this end, the authors are planing to link patients' zipcode to the United States Census Data to retrieve community characteristics (such as median income, educational attainment, unemployment rate, housing status, etc.) for each zipcode and conduct a multilevel regression model to measure the effect of such characteristics on diagnosis of pediatric diseases.

The major limitation of the study is its use of administrative data and ICD codes for case ascertainment, that is sorting people into affected and unaffected groups. while administrative database can provide a large source of low cost observational data on an unselected population, their accuracy depends on accurate diagnoses

by physicians and the validity of the assignment of ICD codes to diagnoses or procedures documented by the physician [28,29]. As a future work of this study the authors will study a more robust and reliable method of case ascertainment such as linking administrative data to other patient information including laboratory and diagnosis test results, vital signs, prescription data, etc.

Declaration of Competing Interest

None.

Acknowledgement

This research is supported in part by Moy Endowed Fund for Collaborative Research at University of Illinois (grant no. MEF2016-6) at Springfield.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.06.020.

References

[1] B.L.W.H.Y. Ma, B. Liu, Integrating classification and association rule mining, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 1998.
[2] A.C. Egberts, R.H. Meyboom, E.P. van Puijenbroek, Use of measures of disproportionality in pharmacovigilance, Drug Saf. 25 (6) (2002) 453–458.
[3] W. Altaf, M. Shahbaz, A. Guergachi, Applications of association rule mining in health informatics: a survey, Artif. Intell. Rev. 47 (3) (2017) 313–340.
[4] S.M. Kang, P.W. Wagacha, Extracting diagnosis patterns in electronic medical records using association rule mining, Int. J. Comput. Appl. 108 (15) (2014).
[5] A. Wright, E.S. Chen, F.L. Maloney, An automated technique for identifying associations between medications, laboratory results and problems, J. Biomed. Inf. 43 (6) (2010) 891–901.
[6] S. Concaro, L. Sacchi, C. Cerra, P. Fratino, R. Bellazzi, Mining healthcare data with temporal association rules: improvements and assessment for a practical use, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2009, pp. 16–25.
[7] R. Harpaz, H.S. Chase, C. Friedman, Mining multi-item drug adverse effect associations in spontaneous reporting systems, in: BMC Bioinformatics, 11, BioMed Central, 2010, p. S7.
[8] M. Rouane-Hacene, Y. Toussaint, P. Valtchev, Mining safety signals in spontaneous reports database using concept analysis, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2009, pp. 285–294.
[9] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin. Pharmacol. Therapeutics 91 (6) (2012) 1010–1021.
[10] V. Ivančević, I. Tušek, J. Tušek, M. Knežević, S. Elheshk, I. Luković, Using association rule mining to identify risk factors for early childhood caries, Comput. Methods Prog. Biomed. 122 (2) (2015) 175–181, doi:10.1016/j.cmpb.2015.07.008.
[11] A.M. Shin, I.H. Lee, G.H. Lee, H.J. Park, H.S. Park, K.I. Yoon, J.J. Lee, Y.N. Kim, Diagnostic analysis of patients with essential hypertension using association rule mining, Healthcare Inf. Res. 16 (2) (2010) 77–81.
[12] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database, Int. J. Endocrinol. Metabolism 13 (2) (2015).
[13] D.G. Lee, K.S. Ryu, M. Bashir, J.-W. Bae, K.H. Ryu, Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction, J. Med. Syst. 37 (2) (2013) 9896.
[14] J. Nahar, K.S. Tickle, A.S. Ali, Y.-P.P. Chen, Significant cancer prevention factor extraction: an association rule discovery approach, J. Med. Syst. 35 (3) (2011) 353–367.
[15] S.H. Park, S.Y. Jang, H. Kim, S.W. Lee, An association rule mining-based framework for understanding lifestyle risk behaviors, PloS One 9 (2) (2014) e88859.
[16] R. Bender, Introduction to the use of regression models in epidemiology, in: Cancer Epidemiology, Springer, 2009, pp. 179–195.
[17] G. Tripepi, K. Jager, F. Dekker, C. Zoccali, Linear and logistic regression analysis, Kidney Int. 73 (7) (2008) 806–810, doi:10.1038/sj.ki.5002787.
[18] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, in: VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, pp. 487–499. http://dl.acm.org/citation.cfm?id=645920.672836.
[19] K. Williams, D. Thomson, I. Seto, D.G. Contopoulos-Ioannidis, J.P. Ioannidis, S. Curtis, E. Constantin, G. Batmanabane, L. Hartling, T. Klassen, Standard 6: age groups for pediatric trials, Pediatrics 129 (Supplement 3) (2012) S153–S160.
[20] C.C. for Medicare, M. Services, 2015 ICD-10-CM and GEMs, 2015. https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs.html.
[21] W. Cooper, icdcoder, 2015. https://github.com/wtcooper/icdcoder.
[22] M. Hahsler, S. Chelluboina, K. Hornik, C. Buchta, The arules r-package ecosystem: analyzing interesting patterns from large transaction datasets, J. Mach. Learn. Res. 12 (2011) 1977–1981. http://jmlr.csail.mit.edu/papers/v12/hahsler11a.html.
[23] S.E. Brossette, A.P. Sprague, J.M. Hardin, K.B. Waites, W.T. Jones, S.A. Moser, Association rules and data mining in hospital infection control and public health surveillance, J. Am. Med. Inf. Assoc. 5 (4) (1998) 373–381.
[24] D.A. Hanauer, N. Ramakrishnan, Modeling temporal relationships in large scale clinical associations, J. Am. Med. Inf. Assoc. 20 (2) (2012) 332–341.
[25] S. Evans, P.C. Waller, S. Davis, Use of proportional reporting ratios (prrs) for signal generation from spontaneous adverse drug reaction reports, Pharmacoepidemiol. Drug Saf. 10 (6) (2001) 483–486.
[26] G. Tripepi, K. Jager, F. Dekker, C. Wanner, C. Zoccali, Measures of effect: relative risks, odds ratios, risk difference, and number needed to treat, Kidney Int. 72 (7) (2007) 789–791, doi:10.1038/sj.ki.5002432.
[27] U.D. of Health, O.o.A.H. Human Services, Trends in teen pregnancy and childbearing, 2016. https://www.hhs.gov/ash/oah/adolescent-development/reproductive-health-and-teen-pregnancy/teen-pregnancy-and-childbearing/trends/index.html.
[28] S. M, R. GE, Finding pure and simple truths with administrative data, JAMA 307 (13) (2012) 1433–1435, doi:10.1001/jama.2012.404. arXiv:/data/journals/jama/23309/jed25024_1433_1435.pdf.
[29] R.E. Hashimoto, E.D. Brodt, A.C. Skelly, J.R. Dettori, Administrative database studies: goldmine or goose chase? Evidence-based Spine-Care J. 5 (2) (2014) 74.