Journal of Biomedical Informatics 53 (2015) 73–80

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

The use of sequential pattern mining to predict next prescribed medications

Aileen P. Wright[a,*], Adam T. Wright[b], Allison B. McCoy[c], Dean F. Sittig[d]

[a] Yale School of Medicine, New Haven, CT, United States
[b] Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[c] Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, United States
[d] The University of Texas School of Biomedical Informatics at Houston and the UT-Memorial Hermann Center for Healthcare Quality & Safety, Houston, TX, United States

ARTICLE INFO

Article history:
Received 13 March 2014
Accepted 8 September 2014
Available online 16 September 2014

Keywords:
Sequential pattern mining
Data mining
Knowledge base
Clinical decision support
Diabetes

ABSTRACT

Background: Therapy for certain medical conditions occurs in a stepwise fashion, where one medication is recommended as initial therapy and other medications follow. Sequential pattern mining is a data mining technique used to identify patterns of ordered events.
Objective: To determine whether sequential pattern mining is effective for identifying temporal relationships between medications and accurately predicting the next medication likely to be prescribed for a patient.
Design: We obtained claims data from Blue Cross Blue Shield of Texas for patients prescribed at least one diabetes medication between 2008 and 2011, and divided these into a training set (90% of patients) and test set (10% of patients). We applied the CSPADE algorithm to mine sequential patterns of diabetes medication prescriptions both at the drug class and generic drug level and ranked them by the support statistic. We then evaluated the accuracy of predictions made for which diabetes medication a patient was likely to be prescribed next.
Results: We identified 161,497 patients who had been prescribed at least one diabetes medication. We were able to mine stepwise patterns of pharmacological therapy that were consistent with guidelines. Within three attempts, we were able to predict the medication prescribed for 90.0% of patients when making predictions by drug class, and for 64.1% when making predictions at the generic drug level. These results were stable under 10-fold cross validation, ranging from 89.1%–90.5% at the drug class level and 63.5–64.9% at the generic drug level. Using 1 or 2 items in the patient's medication history led to more accurate predictions than not using any history, but using the entire history was sometimes worse.
Conclusion: Sequential pattern mining is an effective technique to identify temporal relationships between medications and can be used to predict next steps in a patient's medication regimen. Accurate predictions can be made without using the patient's entire medication history.

1. Introduction

The healthcare system has made considerable headway in the process of transitioning from paper charts to the electronic health record (EHR). This transition has led to the accumulation of vast amounts of data stored in clinical data warehouses which can be used to add to clinical knowledge and guide decision support systems. Data mining is the process of discovering hidden knowledge within a large information repository, and data mining techniques developed for use in retail or other industries can be applied to healthcare [1]. Sequential pattern mining is a data mining technique used to identify patterns of ordered events [2]. In this paper, we use sequential pattern mining to automatically infer temporal relationships between medications, visualize these relationships, and generate rules to predict the next medication likely to be prescribed for a patient.

2. Background

2.1. Stepwise pharmacological therapy

Stepwise pharmacological therapy for management of diseases is common in medicine for progressive conditions such as diabetes mellitus. For example, the American Diabetes Association recommends a treatment algorithm according to progression of disease

[*] Corresponding author. Fax: +1 203 785 6936.
E-mail address: Aileen.Wright@yale.edu (A.P. Wright).

http://dx.doi.org/10.1016/j.jbi.2014.09.003
1532-0464/© 2014 Elsevier Inc. All rights reserved.

in type II diabetes. This algorithm begins with lifestyle interventions and metformin, adds a sulfonylurea if metformin doesn't provide adequate glucose control, then basal insulin, and eventually progresses to using intensive insulin [3]. The algorithm also includes a second tier of less well-validated therapies that instead adds pioglitazone or a GLP-1 agonist to initial lifestyle interventions and metformin, and then may add a sulfonylurea or basal insulin before progressing to more intensive insulin therapy. One can imagine a clinical decision support system that determines where a patient lies within this stepwise algorithm and makes appropriate suggestions to the physician. However, advances in clinical decision support rely on an accurate knowledge base [4]. For example, indication-based prescribing and summarization both rely on a knowledge base of relationships between medications and diagnoses [5–8]. Development and maintenance of an accurate knowledge base by experts is time consuming and expensive. In our past work, we have used frequent item set and association rule mining to infer relationships between medications, laboratory results, and problems [5,9,10]. However, these data mining techniques do not capture temporal information. Little work has been done on the automated development of a knowledge base of temporal relationships between medications, which could be used to guide clinical decision support based around drug regimen changes.

2.2. Sequential pattern mining

Sequential pattern mining is a data mining technique used to identify patterns of ordered events within a database. First introduced in 1995 by Rakesh Agrawal of IBM's Almaden Research Center [11], its original applications were in the retail industry where it can be used to predict that within a certain time period after purchasing a certain book, a customer is likely to purchase its sequel. Applications in medicine were proposed early on [2] and eventually manifested in disease susceptibility prediction [12,13], readmission [14], and pharmacovigilance [15,16].

2.3. SPADE

Identifying all frequent sequential patterns in a transaction database, especially in large databases such as those found in healthcare requires an efficient algorithm to deal with the large search space, and a number of different algorithms have been developed. For example, in a database with 100 different items and sequences up to 5 items long (with item repeats allowed), there would be over a billion potential sequential patterns. In 2001, Zaki described an algorithm called SPADE (Sequential Pattern Discovery using Equivalence classes), which uses a number of strategies to make sequential pattern mining more efficient [17]. Sequential pattern mining typically starts with a transaction database, where each transaction has three fields: the "sequence-id" corresponding to the subject of the sequence (e.g. customer's frequent shopper number or patient's medical record number); the "transaction-time"; and the items associated with the transaction (Table 1). Like previous algorithms, SPADE starts with the horizontal database layout like that seen in Table 1, but it then transforms the dataset into vertical "id-lists" for each item, each consisting of all the sequence-ids and transaction-times where the item is found. Storage of the vertical id-lists allows sequential patterns to be found using intersections of id-lists. For example, the sequential pattern (metformin, insulin) could be found using the intersection of id-lists for the two items. This method minimizes the number of database scans that are required. SPADE also makes use of common prefixes between sequences to decrease the memory requirement. cSPADE is a version of SPADE which incorporates constraints on sequences, such as lengths or time window [18]. It has been applied in protein folding [19], hepatitis classification [20], insider trading detection [21], and satellite image processing [22]. The R package 'arulesSequences' provides an interface to the c++ version of cSPADE [23].

Table 1
Example of transaction database.

Sequence-id	Transaction-time	Items
Patient_1	Aug-2-2008	(metformin, simvastatin, venlafaxine)
Patient_1	Nov-3-2008	(aspirin, glipizide)
Patient_1	July-1-2009	(hydrochlorothiazide, insulin)
Patient_2	Dec-3-2008	(aspirin, azithromycin, metformin)
Patient_2	Aug-5-2009	(insulin)

Recently, Sun et al. [24] used sequential pattern mining to discover common two-item patterns in outpatient data for patients with diabetes; however no study that we know of has used sequential pattern mining to make predictions about next medications likely to be prescribed.

In this paper, we describe the use of cSPADE to identify temporal patterns of medications prescribed for diabetes. We infer temporal relationships from these mined patterns which we visualize in digraphs. We then use the knowledge base of mined patterns to generate rules which predict the next diabetes medication prescribed for a test set of patients.

2.4. Hypothesis

We hypothesize that sequential pattern mining is an effective technique to identify temporal relationships between medications and generate rules that predict which diabetes medication is prescribed next for a patient.

3. Methods

3.1. Definitions

In formal terms, let $I = \{i_1, i_2, \ldots, i_m\}$ be an item set , for example (metformin, simvastatin, venlafaxine). Let sequence s, denoted by $\langle s1,\ s2,\ \ldots,\ sn\rangle$ be a temporally ordered list of item sets, for example ⟨(metformin, simvastatin, venlafaxine), (aspirin, glipizide), (hydrochlorothiazide, insulin)⟩. Let a be another sequence denoted ⟨(metformin), (glipizide) ,(insulin)⟩. Sequence a is called a subsequence of sequence s since (metformin) $\subseteq$ (metformin, simvastatin, venlafaxine) and (aspirin) $\subseteq$ (aspirin, glipizide) and (insulin) $\subseteq$ (hydrochlorothiazide, insulin). A data-sequence is a list of transactions with the same sequence-id (i.e., all transactions belonging to one patient.) The support of sequence a is the fraction of data-sequences which contain a as subsequence. For example, both data-sequences in Table 1 contain the sequence (metformin, insulin) but only 1 out of 2 contains the sequence (metformin, glipizide, insulin) so if this were the complete dataset, the support of (metformin, insulin) would be 1 and the support of (metformin, glipizide, insulin) would be 0.5. The task of sequential pattern mining is to identify frequent sequences, where frequent is defined as having support above a user-defined threshold. In this paper, we will refer to frequent sequences mined from data-sequences as mined sequential patterns, and we will refer to a patient's data-sequence (the temporally-ordered history of all medication prescribed for that patient) as a patient sequence.

3.2. Dataset

We used the dataset described in Parikh et al. [25,26], consisting of inpatient claims data for 6,486,226 members of Blue Cross

Blue Shield (BCBS) of Texas, which is the largest commercial insurance provider in Texas. We extracted a record of all medications prescribed between 2008 and 2011. We included in our study all patients who were prescribed at least one diabetes medication. We then divided the group of patients into a training set (90% of patients) from which to generate rules and a test set (10% of patients) to make predictions for.

3.3. Preparation for data mining

Our first task was to determine the granularity at which we would mine the data (e.g., the drug class, generic drug name only, generic drug and dose, or brand name and dose) since sequential pattern mining can generate a vast number of mined patterns. We decided to mine the data at two different levels of granularity: first at the class level and then at the generic drug level. Our classes were based on World Health Organization Anatomic Therapeutic Chemical (ATC) classes: biguanide, sulfonylurea, insulin, PPAR agonist, DPP-4 inhibitory, GLP-1 agonist, meglitinide, bromocriptine, amylin analog, and alpha-glucosidase inhibitor (Table 2). For mining at the generic drug level, we identified 37 different generic drugs including various 2-ingredient combinations, usually metformin and one additional oral antihyperglycemic ingredient. We treated each ingredient combination as a unique generic drug. We created a mapping table for generic diabetes drugs and another for diabetes drug classes. These tables link each patient's prescription to a generic drug and drug class.

3.4. Sequential pattern mining process

Our sequential pattern mining process is shown in Fig. 1. We performed this process twice; once at the class and once at the generic drug level. We started with a pre-existing database residing in the Microsoft SQL Server relational database management system. We loaded the mapping table into SQL and performed a SQL query to organize the data into the horizontal transaction format required by the algorithm (similar to Table 1). Because we wanted to identify changes in the medication regimen rather than patterns of renewals, we selected only the first prescription of a medication for a patient. For example, if a patient was prescribed metformin, then a glipizide, then had their glipizide renewed, and then was switched to glyburide, their sequence at the class level would be (biguanide, sulfonylurea) and their sequence at the generic drug level would be (metformin, glipizide, glyburide). We used the R package 'arulesSequences' to mine the data for frequent patterns using a support of 1e-10. We chose this very small support to capture a large set of patterns, which we later sorted by the support statistic to identify the most relevant patterns.

3.5. Visualization

We took the mined patterns output by R from the training set at the class level, grouped them by length, and ranked them by support to determine the top sequential patterns. We also used Graphviz [27] to generate a digraph of the 2-item sequences. To make the visualization clearer, we added virtual start and end nodes to each patient's medication sequence before running the sequential pattern mining algorithms to highlight common initiation and termination points, and also pooled data from the training and test sets.

3.6. Evaluation

We set out to explore two research questions: (1) whether sequential pattern mining is useful for predicting changes to a patient's medication regimen and (2) how useful the patient's history of prior medication changes is when making a prediction.

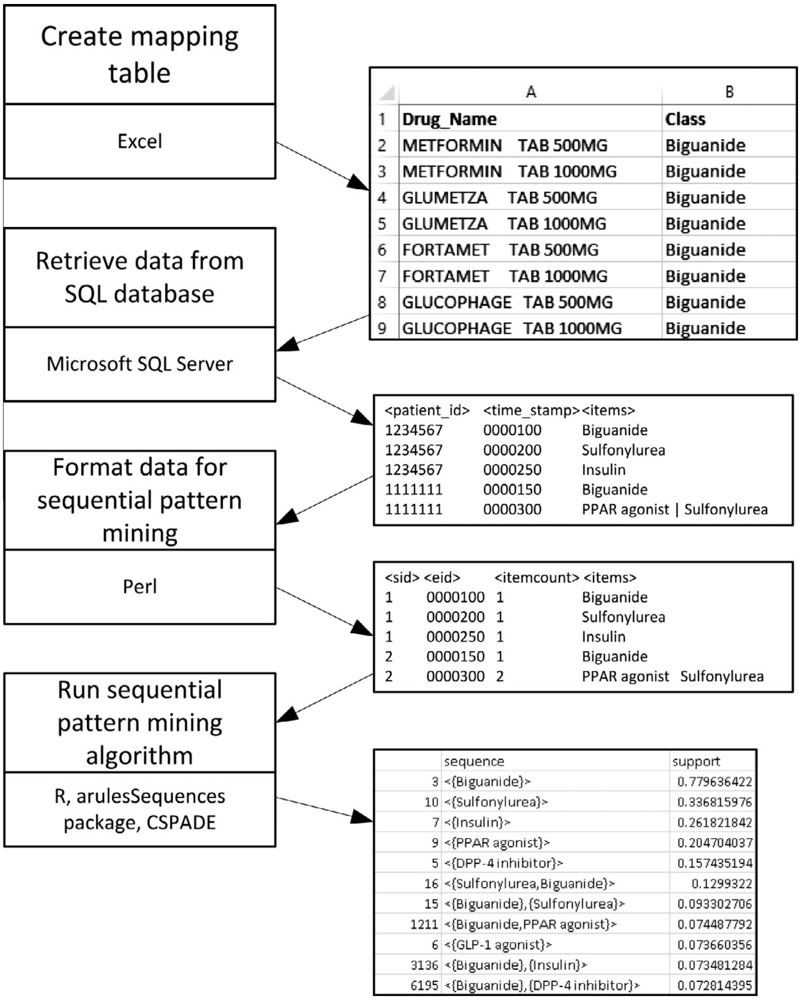
To investigate the first research question, we started with the R output of sequential patterns mined from the training set. To generate prediction rules, we transformed each mined pattern into an antecedent–consequent pair, using the last item of the pattern as the consequent, and all prior items as the antecedent. An example is shown in Fig. 2. Each antecedent–consequent pair served as a rule which, given the antecedent, predicted the consequent.

To prepare the test set for the evaluation, we transformed each patient's sequence into a base stem with an associated "next drug" (Fig. 2). The base stem represents the history of the patient's diabetes medication changes prior to their most recent diabetes prescription. In the special case where a patient had only a single diabetes drug, we used an empty base stem and the patient's only drug as their next drug. To predict the patient's next drug, we looked for rules whose antecedent matched the base stem, ranked them by support, and used the top 5 consequents as predictions for that patient's next drug. We determined how our success changed as we varied how many of the top 5 predictions we included for each patient, with each counting as 1 "prediction attempt". In cases where we were not able to make 5 predictions from a stem, we fell back on predictions made from shorter stems. For example, if we had no rule for a base stem, "Meglitinide → Biguanide → Sulfonylurea → DPP-4 inhibitor", we would use the rule for "Biguanide → Sulfonylurea → DPP-4 inhibitor".

To investigate the second research question, we varied the amount of the base stem used to make our prediction by iteratively truncated each base stem, removing one item at a time from the beginning of the stem. We determined how often the next drug could be correctly predicted within the first three prediction attempts, theorizing that three suggestions would be a reasonable number to display in the EHR. To assess the stability of our results, we used 10-fold cross validation to re-estimate the number of patients for whom the next drug could be correctly predicted within three prediction attempts.

Table 2
Mapping of diabetes drugs to classes.

Drug class	Generic drug
Alpha-glucosidase inhibitor	Acarbose, Miglitol, Voglibose
Amylin analog	Pramlintide
Biguanide	Metformin
Bromocriptine	Bromocriptine
DPP-4 inhibitor	Alogliptin, Anagliptin, Gemigliptin, Linagliptin, Saxagliptin, Sitagliptin, Teneligliptin, Vildagliptin
GLP-1 agonist	Exenatide, Liraglutide, Lixisenatide
Insulin	Insulin lispro, Insulin aspart, Insulin glulisine, Regular insulin, Insulin glargine, Insulin detemir, NPH insulin
Meglitinide	Nateglinide, Repaglinide, Mitiglinide
PPAR agonist	Pioglitazone, Rosiglitazone
Sulfonylurea	Acetohexamide, Carbutamide, Chlorpropamide, Metahexamide, Tolbutamide, Tolazamide, Glyburide, Glibornuride, Glipizide, Gliquidone, Glisoxepide, Glyclopyramide, Glimepiride, Gliclazide


Create mapping table
Excel
Retrieve data from SQL database
Microsoft SQL Server
Format data for sequential pattern mining
Perl
Run sequential pattern mining algorithm
R, arulesSequences package, CSPADE

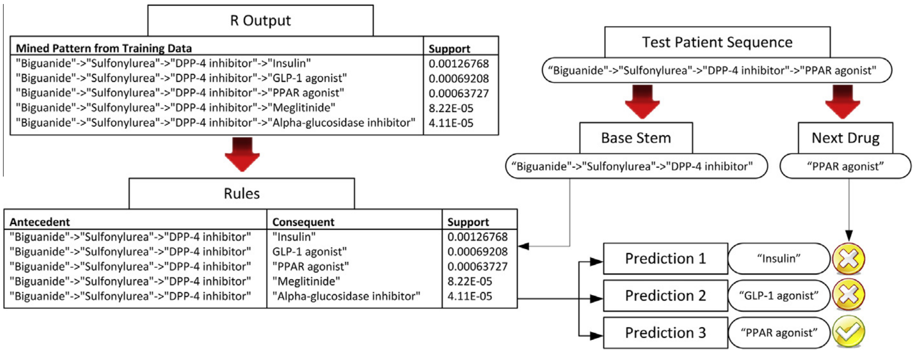
	A	B
1	Drug_Name	Class
2	METFORMIN TAB 500MG	Biguanide
3	METFORMIN TAB 1000MG	Biguanide
4	GLUMETZA TAB 500MG	Biguanide
5	GLUMETZA TAB 1000MG	Biguanide
6	FORTAMET TAB 500MG	Biguanide
7	FORTAMET TAB 1000MG	Biguanide
8	GLUCOPHAGE TAB 500MG	Biguanide
9	GLUCOPHAGE TAB 1000MG	Biguanide

<patient_id>	<time_stamp>	<items>
1234567	0000100	Biguanide
1234567	0000200	Sulfonylurea
1234567	0000250	Insulin
1111111	0000150	Biguanide
1111111	0000300	PPAR agonist | Sulfonylurea

<sid>	<eid>	<itemcount>	<items>
1	0000100	1	Biguanide
1	0000200	1	Sulfonylurea
1	0000250	1	Insulin
2	0000150	1	Biguanide
2	0000300	2	PPAR agonist Sulfonylurea

	sequence	support
3	<{Biguanide}>	0.779636422
10	<{Sulfonylurea}>	0.336815976
7	<{Insulin}>	0.261821842
9	<{PPAR agonist}>	0.204704037
5	<{DPP-4 inhibitor}>	0.157435194
16	<{Sulfonylurea,Biguanide}>	0.1299322
15	<{Biguanide},{Sulfonylurea}>	0.093302706
1211	<{Biguanide,PPAR agonist}>	0.074487792
6	<{GLP-1 agonist}>	0.073660356
3136	<{Biguanide},{Insulin}>	0.073481284
6195	<{Biguanide},{DPP-4 inhibitor}>	0.072814395

Fig. 1. Pipeline of sequential pattern mining process.



Mined Pattern from Training Data	Support
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"->"Insulin"	0.00126768
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"->"GLP-1 agonist"	0.00069208
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"->"PPAR agonist"	0.00063727
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"->"Meglitinide"	8.22E-05
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"->"Alpha-glucosidase inhibitor"	4.11E-05

Antecedent	Consequent	Support
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"	"Insulin"	0.00126768
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"	GLP-1 agonist"	0.00069208
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"	"PPAR agonist"	0.00063727
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"	"Meglitinide"	8.22E-05
"Biguanide"->"Sulfonylurea"->"DPP-4 inhibitor"	"Alpha-glucosidase inhibitor"	4.11E-05

Fig. 2. Evaluation example. Mined patterns from training data are transformed into rules which are used to predict the next drug, given base stems derived from patient sequences in the test set.

We performed our evaluation both at the drug class level and generic drug level. Given the design of our evaluation, we excluded patients with more than one prescription initiated during the same transaction.

4. Results

4.1. Dataset statistics

We identified 161,497 patients prescribed at least one diabetes medication and divided them into a training set ($n = 145,936$) and test set ($n = 16,011$). The most frequent mined sequential patterns for diabetes medication classes are shown in Table 3. Of the top patterns of length 2–6 items, 87% (13/15) begin with a biguanide, 47% (7/15) have a sulfonylurea as the second item, and 67% (10/15) end with insulin. We visualized the top 2-item sequential patterns in a digraph in Fig. 3. Thicker edges between nodes denote a pattern of higher support. The digraph shows a stepwise progression that begins with a biguanide, and either stops there or progresses to a sulfonylurea and/or other non-insulin diabetes medications, and afterwards, may continue on to insulin. Insulin is typically the final step in regimens that have progressed through multiple drugs, as evidenced by the many arrows going from other medications that terminate on insulin.

Table 3
Most frequent mined sequential patterns for diabetes medications, by class.

Sequential pattern	Support
2-item sequential patterns	
Biguanide → Sulfonylurea	0.09308
Biguanide → Insulin	0.07321
Biguanide → DPP-4 inhibitor	0.07261
3-item sequential patterns	
Biguanide → Sulfonylurea → Insulin	0.01089
Biguanide → Sulfonylurea → DPP-4 inhibitor	0.01002
Biguanide → DPP-4 inhibitor → Insulin	0.00858
4-item sequential patterns	
Biguanide → Sulfonylurea → DPP-4 inhibitor → Insulin	0.00127
Biguanide → PPAR agonist → DPP-4 inhibitor → Insulin	0.00123
Sulfonylurea → Biguanide → DPP-4 inhibitor → Insulin	0.00103
5-item sequential patterns	
Biguanide → Sulfonylurea → PPAR agonist → DPP-4 inhibitor → Insulin	0.00013
Biguanide → Sulfonylurea → DPP-4 inhibitor → GLP-1 agonist → Insulin	0.00012
Biguanide → Sulfonylurea → DPP-4 inhibitor → Insulin → GLP-1 agonist	0.00012
6-item sequential patterns	
Biguanide → PPAR agonist → DPP-4 inhibitor → Insulin → GLP-1 agonist → Sulfonylurea	0.00002
Sulfonylurea → Biguanide → DPP-4 inhibitor → PPAR agonist → GLP-1 agonist → Insulin	0.00002
Biguanide → PPAR agonist → Sulfonylurea → DPP-4 inhibitor → GLP-1 agonist → Insulin	0.00002

4.2. Evaluation at drug class level

After excluding patients with more than one drug class initiated during the same transaction, we were left with 121,584 patients in our training set and 11,664 patients in our test set. We were able to correctly guess the next class of diabetes medication prescribed for 6943 (59.5%) of patients using one attempt, and this number increased with successive attempts, so that 10,493 (90.0%) had a correct prediction when 3 attempts were made, and 11,401 (97.7%) when 5 attempts were made (Fig. 4). The majority of patients (67.8%) had a sequence with only one item (i.e., they were prescribed only one class of diabetes medication within the study period), and thus had an empty base stem (Table 4). We were able to make a correct prediction of the medication prescribed for 92.7% of these patients within three attempts. Overall, accuracy ranged from 79.6% to 100.0% when no base stem was used to make a prediction, even for patients with a non-empty base stem. For patients with at least one drug class in their base stem, using a 1-length truncated stem to predict the next drug class was always better than or equal than using the 0-length truncated stem. Using a 2-length stem led to more accurate predictions only for patients with 2 drug classes in their base stem and using 3, 4, or 5-length stems to make a prediction was always worse than or equal to using a 1 or 2-length stem.

4.3. Evaluation at generic drug level

After excluding patients with more than one prescription initiated during the same transaction, we were left with 117,641 patients in our training set and 12,897 patients in our test set. We were able to make a correct prediction for 6637 patients (51.5%) using one attempt, 8270 (64.1%) with three attempts, and 9270 (71.9%) with five attempts (Fig. 4). We were able to correctly predict the next prescribed drug within 3 attempts for 71.4% of the 7778 patients with a 0-length base stem (Table 5). For patients with a base stem containing at least one drug, we were able to correctly predict the next prescribed drug within three attempts 32.6–45.4% of the time without using any portion of their base stem. Accuracy increased when using 1 item from the base stem, and increased further when using 2 items, except in cases where the base stem was 5 items long. Using base stems more than 2 items long was always worse than using 2-length base stems. Overall, the best predictions were made when using 1–2 items from the patient's base stem.

4.4. Cross validation

Under 10-fold cross validation at the drug class level, the percentage of patients with a correct prediction made within 3

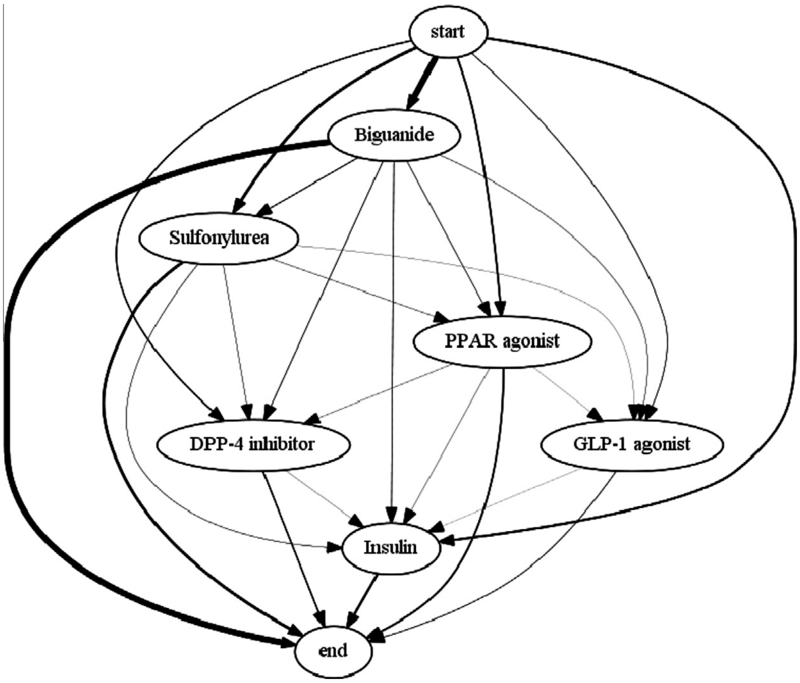


Fig. 3. Digraph of diabetes medications. The most frequent 2-item sequences are shown. Differences in support are represented by edge thickness. For clarity, only the direction between nodes with highest support are shown; reverse directions with lesser support are suppressed.

attempts ranged from 89.1% to 90.5%, with an average of 90.0%. At the generic drug level, this percentage ranged from 63.5% to 64.9%, with an average 64.1%.

5. Discussion

We were able to demonstrate the effectiveness of sequential pattern mining to identify temporal relationships between diabetes medications and reconstruct a stepwise usage pattern that resembles recommendations. Using sequential pattern mining, we were also able predict the next medication likely to be prescribed for a patient.

Table 3 shows a progression of medications that, overall, is consistent with recommendations made by the American Diabetes Association [3], i.e., start with a biguanide, add a sulfonylurea, and progress to basal insulin. In fact, the 3-item mined sequential pattern with the highest support was "Biguanide → Sulfonylurea → Insulin". The addition of a DPP-4 inhibitor or other oral antihyperglycemic to a biguanide was also common but had lower support. This mirrors the American Diabetes Association's algorithm's second tier of less well-validated therapies, which adds pioglitazone or a GLP-1 agonist to initial lifestyle medications and metformin.

While complex, Fig. 3 also reveals clear temporal directions between diabetes drugs consistent with recommendations. For example, there is a clear preference demonstrated for starting a patient on a biguanide before other medications. The second most frequent initial medication is a sulfonylurea. According to guidelines, a sulfonylurea is an acceptable initial medication for patients who have a contraindication to a biguanide (e.g., chronic kidney disease). In the digraph, the nodes for biguanides and sulfonylureas have edges directed towards DPP-4 inhibitors, PPAR-agonists, and GLP-1 agonists, consistent with the American Diabetes Associations description of these as less well-validated therapies. Many pathways converged on insulin, which is consistent with its use in patients with type II diabetes with inadequate glucose control on other medications.

5.1. Evaluation

Using three attempts, we were able to predict the medication prescribed for 90.0% of patients when making predictions by drug class, and for 64.1% when making predictions at the generic drug level. These numbers were very stable under cross validation, with a range of less than 2%. This supports the idea that sequential pattern mining could be useful for predicting which class of medications a prescriber might choose next when treating a progressive disease like diabetes. This prediction could be used to suggest next medications in an EHR when a provider was reviewing medications for a particular problem (Fig. 5). Suggesting more than one medication would increase the accuracy of the prediction, as demonstrated in Fig. 4.

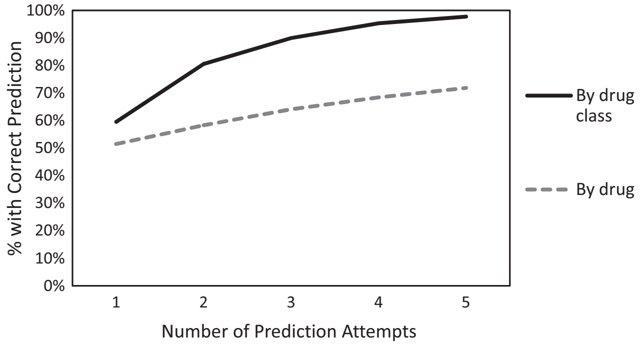


Fig. 4. Accuracy of predictions for next diabetes medication. The percentage of sequences with a correct prediction made is shown for base stems by drug class (dark line) and drug (dotted line).

Performance of the predictions was better at the class level than at the drug level (Fig. 4). Predicting the next drug was a more difficult task since there were 38 possible drugs to predict from, as opposed to 10 possible drug classes. When predicting the next drug class prescribed for a patient whose base stem contained 3 items, there were only 7 remaining drug classes to choose from, so that using 5 attempts it was possible to cover 71% of the remaining drugs. In comparison, when making predictions at the drug rather than drug class level, 5 attempts for a patient with a base stem of 3 items would only cover 9% of all remaining drugs.

When a patient had no drugs in their base stem, i.e., was being started on his or her first diabetes medication, we were able to make a very accurate prediction (92.7% at the drug class level and 71.4% at the generic drug level using three attempts). This accuracy was better than when a patient had a pre-existing medication regimen.

Accuracy was surprisingly high when the patient's base stem was not used to make a prediction, especially when making predictions at the drug class level. Using 1 or 2 items in the patient's base stem led to more accurate predictions, but it was not always better to use the patient's entire base stem, i.e. their entire medication history, and sometimes was worse (Tables 4 and 5). This suggests that doctors may not be conditioning their prescribing behaviors based on the patient's entire medication history. Overall, these results support the idea that sequential pattern mining would be useful for making suggestions in the EHR about which diabetes medication to prescribe next, and accurate predictions could be made without using all of the patient's medication history.

5.2. Applications of the knowledge generated

Sequential pattern mining to identify temporal relationships between medications has applications in clinical decision support when a clinician is reviewing a patient's medication regimen and considering changing it. Guidelines could be made available within the medical record; however, it is expensive and time-consuming to maintain links to the most relevant guidelines in the right place within the medical record. Sequential pattern mining is automated and could guide algorithms that make suggestions when a provider is changing a medication regimen. In Fig. 5 we demonstrate an example for how suggestions for which medication to prescribe next could be incorporated into a patient's problem list. The benefits of having an automatically updated knowledge base of temporal relationships between medications include not only the information provided to the physician (i.e., which medication to prescribe next), but also the convenient placement of the right links to the right medications at the right time, which could save time for users.

The digraphs we created display interesting visual representations of stepwise medical therapy which could be of interest to physicians planning a patient's treatment regimen, as well as patients who wish to understand a typical progression of medications used to treat their condition. These visualizations may also be of use epidemiologically for those who wish to monitor and study the behavior of prescribers at their institution. For example, the typical treatment paths could be mined for different date ranges in order to assess how quickly prescribers are modifying their behavior to adhere to new guidelines.

5.3. Limitations

Since our dataset only captures claims filed between 2008 and 2011, it is likely that some patients were started on other

Table 4
Accuracy of predictions for next diabetes drug class, by sequence base stem and number of previous prescriptions used to make prediction.

No. of drug classes in base stem	No. of patients	Percent with correct prediction made within three attempts, by length of stem utilized					
		0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
0	7906	92.7					
1	2475	79.6	81.1				
2	903	86.4	87.2	88.2			
3	280	95.7	95.7	95.4	94.6		
4	82	92.7	97.6	97.6	93.9	93.9	
5	18	100.0	100.0	100.0	100.0	100.0	100.0

Table 5
Accuracy of predictions of next diabetes drug, by sequence base stem and number of previous prescriptions used to make prediction.

No. of drugs in base stem	No. of patients	Percent with correct prediction made within three attempts, by length of stem utilized					
		0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
0	7778	71.4					
1	3082	45.4	58.8				
2	1230	39.8	46.0	48.2			
3	513	33.9	42.9	46.8	38.8		
4	179	33.5	51.4	52.0	39.1	38.0	
5	86	32.6	52.3	47.7	44.2	40.7	40.7

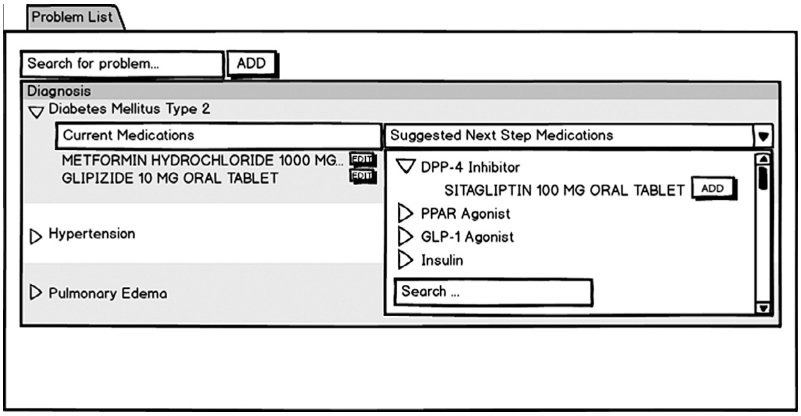


Fig. 5. Medication prediction use case. An example of how predictions for next medication for a disease might be incorporated into the problem list within an electronic medical record.

medications prior to the study period. For example it is possible that patients who appear to have been started initially on insulin were actually initiated on metformin and/or other medications prior to the study period. Further, our dataset only spans 3 years, while progression of diseases such as diabetes can take decades. However, by assembling frequent 2-item sequences into a comprehensive digraph, we were able to visualize temporal relationships between all diabetes medications. In addition, since our database consists of claims data, all patients were insured, and the findings might not apply to uninsured populations.

5.4. Future directions

We aim to explore further methods for using sequential pattern mining to make accurate predictions about which medication a prescriber will choose next, including training rules at the provider level. We also aim to enhance our predictions by using a hybrid model which incorporates not only sequential pattern mining, but also additional profile and clinical data including patient age, gender, comorbidities, and laboratory results into predictions. For example, knowing that a patient has chronic kidney disease would inform which diabetes medication is initially suggested. Once we have further developed this technique, we aim to evaluate the use of sequential pattern mining to make predictions for treatment with antidepressants, antibiotics, chemotherapy, and asthma medications. We also aim to further develop the use of sequential pattern mining to identify and incorporate useful temporal patterns for clinical decision support applications in the EHR.

6. Conclusion

Sequential pattern mining is a useful data mining technique for identifying temporal relationships between medications. From simple two-item sequences, drug regimen pathways can be visualized that fit with guidelines for stepwise drug therapy. These temporal relationships are useful for making predictions about which medication a prescriber is likely to choose next when treating a progressive disease such as diabetes. Future work is necessary to optimize the use of sequential pattern mining to detect temporal relationships among items in the medical record and improve patient care.

References

[1] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE, editors. Medical data mining: knowledge discovery in a clinical data warehouse. In:

Proceedings of the AMIA annual fall symposium. American Medical Informatics Association; 1997.

[2] Srikant R, Agrawal R. Mining sequential patterns: generalizations and performance improvements. Springer; 1996.

[3] Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. Diabetes Care 2009;32(1):193–203.

[4] McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc: JAMIA 2012;19(3):346–52.

[5] Wright A, McCoy A, Henkin S, Flaherty M, Sittig D. Validation of an association rule mining-based method to infer associations between medications and problems. Appl Clin Inform 2013;4(1):100–9.

[6] McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy JA, Butten D, et al. Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. J Am Med Inform Assoc: JAMIA 2012;19(5):713–8.

[7] McCoy AB, Wright A, Laxmisan A, Singh H, Sittig DF. A prototype knowledge base and SMART app to facilitate organization of patient medications by clinical problems. In: AMIA annual symposium proceedings/AMIA symposium AMIA symposium. p. 888–94.

[8] McCoy AB, Wright A, Rogith D, Fathiamini S, Ottenbacher AJ, Sittig DF. Development of a clinician reputation metric to identify appropriate problem-medication pairs in a crowdsourced knowledge base. J Biomed Inform 2013.

[9] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform 2010;43(6):891–901.

[10] Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 2011;18(6):859–67.

[11] Agrawal R, Srikant R. Mining sequential patterns. In: Proc int conf data; 1995. p. 3–14.

[12] Reps J, Garibaldi JM, Aickelin U, Soria D, Gibson JE, Hubbard RB, editors. Discovering sequential patterns in a UK general practice database. Biomedical and health informatics (BHI). In: 2012 IEEE-EMBS international conference on. IEEE; 2012.

[13] Batal I, Valizadegan H, Cooper GF, Hauskrecht M, editors. A pattern mining approach for classifying multivariate temporal data. Bioinformatics and biomedicine (BIBM). In: 2011 IEEE international conference on. IEEE; 2011.

[14] McAullay D, Williams G, Chen J, Jin H, He H, Sparks R et al., editors. A delivery framework for health data mining and analytics. In: Proceedings of the twenty-eighth australasian conference on computer science, vol. 38. Australian Computer Society, Inc.; 2005.

[15] Norén GN, Bate A, Hopstadius J, Star K, Edwards IR, editors. Temporal pattern discovery for trends and transient effects: its application to patient records. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2008.

[16] Jin H, Chen J, He H, Williams GJ, Kelman C, O'Keefe CM. Mining unexpected temporal associations: applications in detecting adverse drug reactions. Inform Technol Biomed IEEE Trans 2008;12(4):488–500.

[17] Zaki MJ. SPADE: an efficient algorithm for mining frequent sequences. Mach Learn 2001;42(1–2):31–60.

[18] Zaki MJ, editor Sequence mining in categorical domains: incorporating constraints. In: Proceedings of the ninth international conference on information and knowledge management. ACM; 2000.

[19] Exarchos TP, Papaloukas C, Lampros C, Fotiadis DI, editors. Protein classification using sequential pattern mining. Engineering in medicine and biology society. In: 2006 EMBS '06 28th annual international conference of the IEEE; 2006 [August 30, 2006–September 3, 2006].

[20] Aseervatham S, Osmani A. Mining short sequential patterns for hepatitis type detection. Proc ECML/PKDD 2005 Discov Challenge; 2005.

[21] Barrientos LA. Insider trading sequential pattern mining (INTRASPAM); 2012.

[22] Julea A, Méger N, Trouvé E, Bolon P, editors. On extracting evolutions from satellite image time series. In: Geoscience and remote sensing symposium, 2008 IGARSS 2008 IEEE international. IEEE; 2008.

[23] Buchta C, Hahsler M, Buchta MC. Package 'arulesSequences'; 2012.

[24] Sun W, Shen W, Li X, Cao F, Ni Y, Liu H et al., editors. Mining information dependency in outpatient encounters for chronic disease care. In: Medinfo 2013: proceedings of the 14th world congress on medical and health. IOS Press; 2013.

[25] Parikh R, editor. Cesarean section rate variation across hospital referral regions in Texas: a claims analysis of privately insured population from 2008–2011. In: 141st APHA annual meeting (November 2–November 6, 2013). APHA; 2013.

[26] Franzini L, Mikhail OI, Skinner JS. McAllen and El Paso revisited: Medicare variations not always reflected in the under-sixty-five population. Health Affairs 2010;29(12):2302–9.

[27] Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G, editors. Graphviz – open source graph drawing tools. Graph Drawing. Springer; 2002.