Journal of Biomedical Informatics xxx (2014) xxx–xxx

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 xxxx

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/

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) ⊆ (metformin, simvastatin, venlafaxine) and (aspirin) ⊆ (aspirin, glipizide) and (insulin) ⊆ (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

Please cite this article in press as: Wright AP et al. The use of sequential pattern mining to predict next prescribed medications. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.09.003