Mining association rules and sequential patterns from electronic prescription databases

Daniel Botas
Instituto Superior Técnico, Universidade de Lisboa
daniel.botas@tecnico.ulisboa.pt

Abstract

Over the years many scientific studies have been published in Medicine to evaluate, understand and predict the effects of introducing new medications. However, those studies draw conclusions from small samples, due to the difficulty and cost of retrieving large quantities of data through questionnaires. Thanks to the growing trend in prescription process automation, large amounts of medical data are now stored in databases that can later be explored to discover potentially useful information. Electronic Prescription data can be analyzed to improve prevention, diagnosis and treatment of diseases, optimize resources, and promote patient safety. This dissertation presents a methodology to discover association rules and frequent sequences in databases of electronic prescriptions using the Apriori and PrefixSpan algorithms. The methodology was used to characterise the Portuguese population prescribed with anticoagulants. This study enabled (a) an assessment of the adoption of novel oral anticoagulants, including the identification of predictive factors associated with discontinuation or changes of prescribed medication, (b) discovery of causal association rules between medications, and (c) characterization of frequent patterns associated with the consumption of anticoagulants. The main conclusion of this work is that data mining techniques can be applied to electronic prescription databases to extract knowledge which can latter support decision-making in public health.

Keywords: Anticoagulant prescriptions analysis, Electronic prescriptions mining, Frequent and sequential patterns, Data mining

1. Introduction

For the purpose of optimization of process management, countries worldwide are adopting centralized management systems to reduce fiscal fraud and optimize costs. In Europe, and in particular in Portugal, integrated systems supporting electronic prescriptions are being used as a measure to improve safety, quality, efficiency and cost-effectiveness in healthcare. As a result, large scale prescription databases covering entire nations are now ready to be analyzed for secondary purposes.

While data mining techniques have been applied successfully in other types of databases [1, 2, 3, 4], their application over electronic prescription databases has not been greatly explored. However, data mining algorithms could be applied to prescription databases to uncover interesting patterns (e.g., co-occurrences between different medications, sequences of medications appearing frequently and corresponding to common treatment regimes). In this study, we explore a dataset containing the electronic prescriptions from the Portuguese national healthcare system comprising prescription data between the years 2013 and 2016, using association rules (i.e., there is a list of antecedent medications and a list of consequent medications appearing frequently and corresponding to common treatment regimes). In this study, we explore a dataset containing the electronic prescriptions from the Portuguese national healthcare system comprising prescription data between the years 2013 and 2016, using association rules (i.e., there is a list of antecedent medications and a list of consequent medications), causal rules (i.e., similar to association rules but the association is stronger between the antecedent and consequent, namely the fact that taking the antecedent implies that the consequent is also taken) and frequent sequences between prescribed medications. The most important contributions of this work can be summarized as follows:

- We introduce data pre-processing strategies for transforming a set of prescription records into a suitable database for data mining tasks, envisioning the discovery and exploration of relationships between prescribed substances, using the concepts of Defined Daily Dosage (DDD) to define treatment regimes.
- We describe the application of association rules to discover causal rules between medications from electronic prescription databases, a task that remains largely unexplored in the context of biomedical informatics.
- We identified suitable techniques to discover parallel clinical pathways in treatment regimes.
- From the application of the Apriori algorithm, it was possible to observe that the rules associated with the male patients have a greater lift when compared with the female ones. Also, it was interesting to note that not only the female top rules contain 33% more medications than the male’s, but also that the New Oral Anticoagulant (NOAC) Rivaroxaban already appears in the top rules. The comparison between the association rules from Lisbon and Porto, the two biggest districts in terms of population, showed that Porto produces association rules with a much larger lift.

• From the application of the PrefixSpan algorithm, it was possible to observe that in age group 0-44 Warfarin appears linked, as expected, to medications used to treat hypertension and cholesterol, but also to less expected treatments like arrhythmia. Even more surprising was the connection found, in age group 65-74, between NOAC Rivaroxaban and medications used in insomnia.

The rest of this report is organized as follows: Section 2 presents the data mining procedures, including data pre-processing strategies, and the algorithms for mining association rules and sequential patterns. Section 3 presents a case study leveraging the proposed methodology, with an initial data characterization, followed by the results achieved with the proposed methodology. Section 4 provides an overview on previous studies addressing data mining from healthcare databases, and specifically focusing on prescription databases. Finally, Section 5 provides the conclusions obtained from this work, together with directions for future research.

2. Methods

This section presents the methods that were proposed to discover interesting patterns in medical prescription databases using data mining techniques. In subsection 2.1 we describe in detail the sequence of transformations applied to the data to be suitable for data mining algorithms. Then, in subsection 2.2 we explained the application of the data mining algorithms in this work, including the evaluation metrics used. The process to discover useful knowledge from the databases consisted in five phases:

1. Selection: This phase consists on creating a target dataset, or focusing on a subset of variables or data samples. In this study, we focused on the prescriptions from Portuguese patients that had at least one prescription of anticoagulants between the years of 2015 and 2016.

2. Pre-processing: This phase consists on the target data cleaning and pre-processing in order to obtain consistent data. Due to the longitudinal nature of this study, special care was taken to carefully examine and consolidate the attribute fields.

3. Transformation: This phase consists on the transformation of the data using dimensionality reduction or transformation methods. We used external information about medication treatments, in this case the concept of DDD, to transform a database of electronic prescriptions into one that indicates how patients are being medicated throughout time, e.g. in terms of the different active substances and patient treatment regimes.

4. Data Mining: This phase consists on the searching for patterns of interest in a particular representational form, depending on the Data Mining objective. The algorithms Apriori and PrefixSpan were applied to discover association rules and frequent sequences between medications.

5. Interpretation: This phase consists on the interpretation and evaluation of the mined patterns. Based on the framework Rule Changing + Relevance Feedback, we used evaluation measures Support, Confidence and Lift, combined with concepts of Unexpectedness, Novelty and Cohort Studies, to evaluate and rank the results.

2.1. Data Preparation

Mining interesting patterns requires first transforming a database containing a chronologically ordered sequence of prescriptions into one that indicates how patients are being medicated throughout time, e.g. in terms of the different active substances and patient treatment regimes.

In an electronic prescription databases one is expected to find fields regarding patients, including a unique identifier for the patient, as well as descriptive attributes like age, gender, and location (e.g., administrative district). These last fields, in particular, must be carefully examined and consolidated, since in the context of longitudinal studies patients can change throughout time. Expected fields regarding active substances include a unique name for the substance, the type of administration, prescription dosage, package size, and number of packages. Finally, attributes for characterizing the context of a prescription include a date, an identifier for the doctor making the prescription, and the doctor’s medical speciality.

None of the fields above contains an indication on how many days a given prescription can maintain a person under treatment. To estimate this, we can leverage the concept of Defined Daily Dosage (DDD), which corresponds to the assumed average maintenance dose per day for a medicine, considering its main indication in adults. Although the DDD does not necessarily reflect the prescribed daily dosage, it gives a rough estimate of consumption, providing a fixed unit of measurement independent of price, currencies, package size, and strength. Considering the DDD, the duration, in days, of a treatment can be estimated by:

\[
\text{Nr Days} = \frac{\text{Nr Packages} \times \text{Nr Pills per Package} \times \text{Dosage}}{\text{DDD}} \tag{1}
\]

This information can be used to reconstruct the time intervals for which each patient was subjected to a certain treatment.
our study, to determine the correct DDD of a given prescription, we made a manual mapping between the active substance, the therapeutic group, and the medical speciality, using information collected from the World Health Organization website, combined with doctor Paulo Nicola medical expertise. Using this information, the original records were augmented with attributes for the therapeutic group, number of days associated to the treatment, and the expected end day for the treatment.

Patient-Day transactions are formed by grouping records by patient and day. Each of these transactions correspond to all the active substances prescribed to a particular patient on a given day, according to the estimates provided by Equation 1. In the example shown in Figure 1, we can see that the patient identified by E3A9629 took Paracetamol continuously from 01/07/2016 until 04/08/2016. He also started taking Bisoprolol and Rivaroxaban from 01/08/2016. The corresponding Patient-Day transactions for this example shows that on 31/07/2016 this patient was only taking Paracetamol, but on the next day he was taking Paracetamol, Bisoprolol and Rivaroxaban at the same time. In cases where a patient has multiple prescriptions of the same medication with overlapping dates between them, can be combined into a single treatment period by increasing the end date of the first prescription with the treatment days of the subsequent overlapping prescriptions.

Since transactions can contain fields with missing or contradictory values (e.g. multiple gender changes, missing age, missing speciality), I attributed age, gender, and district to the most frequent values inside the transaction (in cases where no frequent value was available inside transaction, it was used the most frequent value associated with the patient).

To identify patient treatment regimes, consecutive records of the same active substances belonging to the same patient from Patient-Day transactions are truncated to only one record (most recent date) meaning one treatment. Now, Patient-Prescription Regime transactions are formed by grouping all the previous records by patient id. Each transaction corresponds to a period of continuity in terms of a patient’s treatment regime. These transactions enable the assessment on treatment changes as an indicator of a patient adherence to a different treatment regime, including possible causes. To better identify the treatment regimes, I assumed that two regimes with equal active substances constitute one treatment if the time interval between them is less than 2 days, otherwise they are treated as two separate regimes. In the example shown in Figure 1, consecutive transactions with three substances (i.e., of Paracetamol, Bisoprolol and Rivaroxaban) were grouped into a single transaction Patient-Day-Prescription Regime representing changes in treatments.

To identify the sequential patterns, the input consisted of the prescriptions from each patient, chronologically ordered. In an optimized version, the entry sequences contained all the prescriptions with less than 30 days between them, i.e patients with very spaced prescriptions are represented by several entry sequences, instead of only one.

For both association rule and frequent item-set mining algorithms, different sub data-sets were created to facilitate prescription comparisons between age groups, gender and district.

2.2. Data Mining

This section provides an explanation on the algorithms used to perform data mining over the generated datasets, including also metrics that are used in the assessment of the extracted rules.

2.2.1. Frequent Itemsets and Associations Between Frequent Itemsets

To discover associations rules between prescribed medications, it is commonly used a data mining technique named Association Rule Mining (ARM) which focuses on finding all large item-sets (i.e., collections of items) and association rules between item-sets that satisfy both syntactic and support constraints. Notice that association rules encode strong co-occurrence patterns, although these do not typically imply causality between items. To define association rules, let us assume \( I = \{i_1, i_2, \ldots, i_n\} \) to be a set of items (i.e., each item corresponds to a prescribed medication) and \( T \) to be a set of transactions called a database. A transaction \( t \) (i.e., each transaction corresponds to all the prescribed medications associated with a particular patient) from \( T \) can be represented by a unique identifier and a binary vector that represents the occurrence/absence of a subset of items in \( I \). With these definitions, we can say that an association rule is an implication:

\[
X \Rightarrow I_j
\]

In the previous expression, \( X \) is a subset of \( I \) and \( I_j \) is another subset of \( I \) that is absent in \( X \). Restrictions on the items that appear in Expression 2 are syntactic constraints, while association strength restrictions are usually expressed by evaluation metrics such as support, confidence, or lift.

Apriori is an interactive breadth-first search algorithm that uses a generate-and-test strategy to mine frequent item-sets for association rules. It is based on the principle that if an item-set is frequent, then all of its subsets must also be frequent. Leveraging on this principle, Apriori can prune candidate item-sets with infrequent subsets without having to count their support (i.e., the fraction of transactions in \( T \) that contain the item-set). This holds true due to the anti-monotone property of the support for an item-set, i.e., the fraction of transactions on a database \( T \) that contain all the items in the item-set under analysis:

\[
\forall X, Y : (X \subseteq Y) \Rightarrow s(Y) \leq s(X)
\]

In the previous expression, \( X \) and \( Y \) represent item-sets and \( s(X) \) represents the support associated with item-set \( X \). Expression 3 denotes that the support of an item-set never exceeds the support of its subsets.

To make an assessment on the strength of an association rule, common metrics include support, confidence, and with lift, with this last particular metric showing to what extent the antecedent and consequent of a rule are dependent on one another. 

\[\text{http://www.whocc.no/ddd/}\]
Let $X \Rightarrow Y$ be a rule where $X$ and $Y$ are disjoint item-sets (i.e. $X \cap Y = \emptyset$). Support shows how often a rule, with respect to a set of transactions $T$, can be applied to a data-set. Rules that have low support typically occur by fortuity and often are uninteresting.

$$\text{supp}(X \Rightarrow Y) = \frac{|\{t \in T : X \subseteq t\}|}{|T|} \quad (4)$$

Confidence shows how often the association rule has been found to be true, i.e., the reliability of the association made by the rule. Confidence thus estimates the rule’s conditional probability. The higher the confidence value, the more often this set of items is associated together.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (5)$$

Lift is a symmetric measure with respect to the antecedent and consequent of a rule, that measures the co-occurrence (not implication) in order to retrieve rare important rules, typically pruned by user-defined support and confidence thresholds [8].

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} \quad (6)$$

A lift value greater than 1 indicates that the occurrence of the rule body has a positive effect on the occurrence of the rule head. If the value is smaller than 1, this means that the occurrence of the rule body has a negative effect on the occurrence of the rule head. If the value is 1, the occurrence of the rule body has almost no effect on the occurrence of the rule head.

It is important to emphasize that associations do not necessarily imply causality, though the reverse is true. This means that cause-effect relationships explaining how events have developed, or the ability to make forecasts, cannot be directly deduced from association rules. To identify causal relations, a previous study proposed an odds ratio measure inspired on retrospective cohort studies [9]. These studies select individuals who have been exposed or not to a suspected risk factor, but are alike within many other aspects. Notable differences between the results of the two cohorts indicate a possible causal relationship between exposure and response variables. Leveraging on this idea, causal rules can also be defined in a similar manner.

Let us consider a data-set $T$, and the association rule $P \Rightarrow Z$ as an hypothesis. Let $P$ be the exposure and $Z$ the response variables, with $C$ representing the set of control variables. The process begins by choosing a record containing $P$, and then another record not containing $P$, while $Z$ is blinded, and both records have matched values for $C$. Then, both of the records on the matched pair are removed from $T$, and attributed to the corresponding group. The process repeats until no more matched pairs can be found. The result is a fair data-set, corresponding to the maximal sub data-set of $T$ that contains only matched record pairs.

There are four possibilities for a matched pair: both records contain $Z$ ($n_{11}$), neither contain $Z$ ($n_{22}$), the record belonging to the exposure group containing $Z$ and non-exposure does not ($n_{12}$), and vice versa ($n_{21}$), as shown in Table 1. With basis on the group contingency table, the odds ratio of a rule $P \Rightarrow Z$ on a fair data-set $T$ can be defined as:

$$\text{OddsRatio}_T(P \Rightarrow Z) = \frac{n_{12}}{n_{21}} \quad (7)$$

When the odds ratio of an association rule on its fair data-set is significantly greater than 1, this means that a change of the response variable is a consequence of the exposure variable, and thus indicative of a causal rule.

### 2.2.2. Sequential Patterns and Associations Between Sequential Patterns

To discover frequent sequences in prescribed medications, a data mining technique called Sequential Pattern Mining (SPM) can be used. Agrawal and Srikant [10] described it as follows: given a set of sequences, where each sequence consists of a list of elements, and given a user-specified min-support threshold, SPM concerns with finding all of the frequent sub sequences, i.e., the sub sequences whose occurrence frequency in the set of sequences is no less than min-support.

Although Apriori can be used for SPM [11], the complexity associated with the generation of rules leads to the appearance of new and more efficient algorithms. Prefix-Projected Sequential Pattern Mining, also known as PrefixSpan [12], is one of them that are capable of dealing with very large databases. PrefixSpan is an example of a pattern growth algorithm, that examines only the prefix sub-sequences (i.e., the first prescribed medications) and projects only their corresponding postfix subsequences (i.e., the subsequent prescribed medications) into projected databases in order to make the databases for the next pass smaller. There are three projection methods available for PrefixSpan: level-by-level where there is no candidate generation step and the sequential database is divided into different partitions according the number of length-1 sequence, i.e. each partition is the projection of the sequential database that take the corresponding length-1 sequences as prefix; bi-level that uses triangle matrix instead of a projected database to reduce the number and size of projected databases; pseudo-projection that based on evidences that postfixes of a sequence often appear repeatedly in recursive projected databases, instead of constructing a physical projection by collecting all the postfixes, uses pointers to refer to the sequences in the database as a pseudo-projection. However, the main limitation with this method is that the database must be stored into the main memory.

In each projected database, sequential patterns are grown by exploring only frequent local patterns. Thus, the major cost in PrefixSpan is to build projected databases (in the worst case,PrefixSpan constructs a projected database for each sequential pattern). The sequences generated by PrefixSpan can

<table>
<thead>
<tr>
<th>$P = 0$</th>
<th>$P = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z$</td>
<td>$\neg Z$</td>
</tr>
<tr>
<td>$\neg Z$</td>
<td>$n_{21}$</td>
</tr>
</tbody>
</table>

Table 1: Contingency table
then be evaluated by their support, absolute frequency, and by maximum pattern length.

To discover clinical pathways amongst prescribed medications, a technique a technique named Post Sequential Pattern Mining [13] (PSPM) can be used to discover complex structural patterns hidden behind sequences, in order to represent relations between sequential patterns (i.e., sequences of medications representing treatments). This method leverages on traditional SPM techniques to generate sequential patterns, and then continues processing them to discover hidden Structural Relation Patterns (SRP). The main idea behind this method is to search for sequential patterns supported by data sequences, which can then be used to determine SRP patterns, and subsequently to determine the corresponding maximal set. Since sequential patterns supported by data sequences can be viewed as a transaction, the problem of mining SRP, in particular concurrent patterns satisfying a minimum confidence, is similar to mining frequent item-sets under minimum support.

One particular case of SRP are concurrent patterns, which corresponds to parallel clinical pathways in this work. Concurrency can be defined as the fraction of data sequences that contain all of the sequential patterns [14]. Let us assume that SDB is a sequence database with unique values for each record. Assume also that \( \{s_{p_1}, s_{p_2}, ..., s_{p_m}\} \) is the set of \( m \) sequential patterns mined under a given min-support threshold, and that they are not contained in each other. Then, a pattern concurrency can be defined as follows:

\[
\text{conc}(s_{p_1}, s_{p_2}, ..., s_{p_i}) = \frac{|\{C \in \text{SDB} : \text{sp}_i \subseteq C\}|}{|\text{SDB}|}
\]

In the previous equation, \( s_{p_i} \subseteq C \) indicates that the sequential pattern \( s_{p_i} \) is contained in data sequence \( C \). Concurrent Sequential Patterns (CSPs) are sequential patterns whose concurrency is above a min-confidence threshold. CSPs are represented by ConSP\(_k\) = \( \{s_{p_1} + s_{p_2} + ... + s_{p_k}\} \) where \( k \) is the number of sequences and + denotes the concurrent relationship.

To mine these patterns, we begin by identifying the sequential patterns who have support in the sequential patterns occurring together. Notice that ConSP\(_k\) assures that the patterns occur above a certain threshold, although it does present its minimal representation, as further relations have not yet been explored. These patterns can be viewed as transactions and the problem of finding CSPs satisfying min-confidence can be solved using techniques to mine frequent item-sets satisfying min-support (e.g., the aforementioned Apriori algorithm). The resulting patterns must then be simplified to ensure they are not contained in any other concurrent pattern. The simplified set can be obtained by deleting concurrent patterns which are contained by other concurrent patterns, and/or by deleting sequential patterns which are contained by other sequential patterns, when mining for frequent item-sets.

3. Results

In this section we present the results obtained by the execution of the methods described in the previous section. First, we present a statistical characterization of the dataset supporting the case study. Then, the obtained results are presented and discovered in detail.

3.1. Dataset Characterization

An initial statistical characterization from the Portuguese medical prescriptions database regarding years 2015 and 2016 has been made, focusing on prescriptions for common anticoagulants (i.e., chemical substances used in therapy for thrombotic disorders, that prevent or reduce coagulation of blood, thus prolonging the clotting time). In a total of 4192007 medical prescriptions, from 30396 different medical doctors and involving 125196 different patients, approximately 14.1% correspond to anticoagulants with 28.3% of those corresponding to the prescription of Warfarin, the most common anticoagulant and also the most common medication being prescribed overall in Portugal with Rivaroxaban (i.e., the most popular NOAC) in the second place. Interestingly, the number of medical doctors prescribing Rivaroxaban is in fact slightly larger than the number of doctors prescribing Warfarin, although the number of prescriptions for Warfarin is much larger. The monthly fluctuations shown in Figure 2 reflect the changes in the total number of monthly prescriptions and number of distinct doctors prescribing anticoagulants, and the same figure also shows that the relative number of prescriptions for the different types of anticoagulants remains relatively stable (although the percentage of NOACs versus traditional anticoagulants seems to be increasing). As expected, anticoagulant prescriptions are much more common in patients over 75 years of age.

3.2. Association Rules

We used association rules to characterize the relationships between medications prescribed (antecedent) and the next one to be prescribed (consequent). To visualize the discovered association rules, we used nodes and arrows to connect the antecedent to the consequent of a rule. The size and color of a node provide a visual representation of the lift and odds ratio of the mined rules.
value of a rule. Rules that have large nodes represent potentially interesting rules, while those with darker colors indicate causal rules. For example, looking at figure 3 representing the top 10 rules (ordered by lift and odds ratio in this order) in Portugal, the rule Furosemida, Varfarina $\Rightarrow$ Espironolactona has a high lift and odds ratio, implying that Espironolactona is specifically used with Furosemida and Varfarina, and a good candidate for a causal rule. However, it is interesting to note that Varfarina, Espironolactona $\Rightarrow$ Furosemida has a smaller lift and odds ratio. This implies that Furosemida is frequently used with other medications and can be considered as a commonly used medication.

Observing the top rules by gender in Figure 4, it is possible to see that the rules associated with the male patients have a greater lift when compared with the female ones. Also, it is interesting to see that not only the female top rules contain 33% more medications than the male’s, but also that the NOAC Rivaroxaban, already appears in in top rules.

Looking at the top rules by age groups in Figure 5, it is possible to see that in the age group 0-44, there are evidences that support that a causal rule between the taking of Escitalopram, Nebivolol and Candesartam implies the taking of Warfarin, the only anticoagulant present in this group. In group 45-54 there are two anticoagulants, the older and cheaper Warfarin and the more expensive NOAC Rivaroxaban. Also in this group, there is a possible causal rule between the taking of Nebivolol together with Alprazolam and the consequent taking of Rivaroxaban. Another type of comparison was made using the districts where the prescriptions were made. Comparing the Lisbon and Porto districts (biggest population centers in Portugal with sim-
Similarly sized population) in Figure 6, we can observe that top rules in Porto have a significantly larger lift than those in Lisbon. Also in the same figure, the top rules in the coastal district of Coimbra and the inner district of Guarda are also shown. However due to the low population in inner districts, and thus fewer prescriptions, the rules discovered in these types of districts typically have a bias associated.

3.3. Sequential Patterns

To discover statistically relevant patterns present in Patient-Day transactions, we used PrefixSpan algorithm to identify frequent patterns in the transactions. To filter the results we used a minimum support of 5% and a minimum length of 3 medications.

To visualize the discovered patterns we used nodes to represent the medication and arrows (in different colors) to connect the sequences. For example, in figure 7 there are two patterns (represented by green and orange arrows) that begin with Atorvastatin, then in one pattern (orange) the next medication is Candesartan, while the other pattern (green) is followed by Propafenone, and both frequent patterns end in Warfarin.

Observing the frequent sequences from the dataset corresponding to age group 0-44 in figure 7 we can observe that old type anticoagulant Warfarin appears linked, as expected, to medications used to treat hypertension (Candesartan) and cholesterol (Atorvastatin) but also to less expected treatments like arrhythmia (Propafenone).

Looking at the frequent sequences from the dataset corresponding to age group 65-74 in figure 8 it is interesting to note that the NOAC named Rivaroxaban appears in an older age group linked with the treatment of cholesterol (Sinvastatine), and although not so expected, also with the treatment of insomnia (Zolpidem) representing more than 5.5% of the transactions in this group.

Our data-set corresponding to age group 45-54 did not return any frequent sequence with minimum support of 5% and with at least length 3. With regard to the other two data-sets (55-64 and 75+), they could not be processed in time to be evaluated and presented in this report due to the expansion of several records containing large number of items (≥ 10).

4. Related Work

The transition from paper records to Electronic Medical Record (EMRs), together with the adoption of electronic prescriptions, helped foster experiments with a number of data mining techniques [15, 16, 17, 18, 19, 20, 21, 22]. For instance, Wright et al. [15] explored how Association Rule Mining (ARM) techniques can be applied over EMRs to discover clinically accurate and meaningful associations between medications and diseases. The authors extended the Apriori [5] algorithm with an iterative transitive inference to prune spurious candidate associations, assessing rules with multiple statistics of interestingness. The obtained results support the idea that associations generated by ARM are not only accurate and take less time do generate than using human knowledge, but they are also more readily characterized and validated, thus generating a volume of rules that is significantly larger than those in existing expert-curated knowledge databases. The authors also observed that no single statistical measure for rule assessment is the best,
and instead evaluation metrics should be carefully chosen depending on the data-set characteristics and applications.

In a related study, Wright et al. [16] studied the feasibility of identifying temporal relationships between medications, with the aim of predicting the next medication likely to be prescribed using Sequential Pattern Mining [12] (SPM). The process begins with the selection of a desired level of granularity (e.g., analyse sequences of active substances or sequences of therapeutic groups) for the events. Then, transactions can be created either by simply using fields such as patient id, transaction time and the event sequences associated with units of analysis, or more refined approaches such as the one proposed by Tôth et al. [17], which represents transactions based on the desired level of detail (e.g., level 1: surgery, level 2: classical surgery or endoscopic surgery) without the loss of valuable information. The creation of transactions was achieved through the usage of a coding system combined with event aggregation reflecting professional medical practices (e.g., aggregating repeating events so that only the first one appears, choosing only the most relevant event from multiple similar same day events). The formed transactions are then passed the SPADE [23] sequential pattern mining algorithm, which extracts antecedent-consequent pairs. For each patient, the corresponding set of antecedents is seen as a base stem, and the consequents correspond to possible next medications. The prediction was achieved by looking at pairs that matched the base stem, and then ranking them by support.

The aforementioned prediction strategy, based on frequency, can be biased towards personal prescription strategies in situations where a small number of doctors are responsible for a large proportion of prescriptions. To reduce this bias, Morita et al. [18] proposed a score for the discovered sequences called DF-Score, combining the support of a sequence and a score for unbiased usage. The unbiased usage score is calculated as the proportion between the rate of doctors who prescribed some sequence and the unbiased estimator of variance, representing the usage frequency dispersion for that particular sequence, for each doctor.

Perer et al. [20] observed that in EMRs, many clinical events can occur and/or be registered simultaneously as a single event (e.g., prescription together with blood pressure evaluation), creating an explosion of possibilities that can hinder frequent pattern mining algorithms. To address this problem, the authors proposed to find frequent clinical event packages (CEPs), that are subsets of Same Day Concurrent Events (SDCEs). Treating each SDCE in every patient trace as a transaction, the problem of detecting those CEPs is equivalent to the problem of frequent item-set mining, with each identified CEP acting as a super event.

Also in the context of mining EMRs, Huang et al. [21] explored the concept of comorbidities and how they influence the adoption of essential treatments in Clinical Pathways (CPs). Using a probabilistic statistical model, diagnose labels and consequent treatment events were linked together in order to unveil the latent associations between diagnosis labels and treatments, and compute the contribution of comorbidities on treatment adoptions in CPs. Results showed that the proposed model can discover not only meaningful latent treatment patterns exhibiting comorbidities traits, but also implicit changes of treatments of first-diagnosis due to the incorporation of typical comorbidities.

Noting that comorbidities need to be studied in the context where they occur, Boytcheva et al. [22] proposed a model in which clinical events are enriched with semantic attributes (e.g., patient demographics, age, gender, localization, etc). This way, risk groups of patients with a predisposition to develop socially-significant disorders can be discovered using retrospective analysis of population data, by filtering events with common properties and similar significance.

5. Conclusions and Future Work

With our work we have defined the methods to analyze entire treatment regimes from entire populations that can be easily applied to other e-Prescription databases. To that end, we defined a novel methodology to transform a set of medical records into a suitable database for data mining tasks, envisioning the discovery and exploration of relationships between prescribed medications.

We have contributed with a thorough and detailed research on the state-of-the-art, studying the existing data mining techniques and evaluation methods, in particular those focused on mining prescriptions and other health record databases. Another important contribution of our work was the development of adequate visualization techniques to present the information the reader in a clear and concise way.

Our results showed that regarding association rules, the rules associated with the male patients have a greater lift when compared with the female ones. Also, it was interesting to note that not only the female top rules contain 33% more medications than the males, but also that the NOAC Rivaroxaban already appears in the top rules. The comparison between the association rules from Lisbon and Porto, the two biggest districts in terms of population, showed that Porto produces AR with much larger lift. With regards to the results for sequential patterns, it was possible to observe that in age group 0-44 Warfarin appears linked, as expected to medications used to treat hypertension and cholesterol, but also to less expected treatments like arrhythmia. Even more surprising was the connection found, in age group 65-74, between NOAC Rivaroxaban and medications used in insomnia, accounting for more than 5.5% of the transactions in this group.

Taking into account the focus of our study, related future works on this subject may involve the definition of new types of transactions (e.g., Medic, Speciality, Prescription) to provide different perspectives over the same data. I also recommend using the hierarchical classification of active substances to provide different levels of detail. Another recommendation I strongly suggest is the use of the Post Sequential Pattern Mining algorithm to discover complex structural patterns hidden behind sequences, in order to represent relations between sequential patterns.
6. Bibliography


