Journal of the Midwest Association for Information Systems (JMW AIS)

Volume 2017

Issue 2 Health Information Systems - Integration in the
Global Healthcare Ecosystem, Special Issue Editor:
Ciara Heavin, University College Cork, Cork, Ireland

Article 2

2017

Analysis of a Sequence of Events in Healthcare

Pankush Kalgotra

Clark University, pkalgotra@clarku.edu

Ramesh Sharda

Oklahoma State University, ramesh.sharda@okstate.edu

William D. Paiva

Oklahoma State University, wpaiva@okstate.edu

Follow this and additional works at: <http://aisel.aisnet.org/jmwais>

Recommended Citation

Kalgotra, Pankush; Sharda, Ramesh; and Paiva, William D. (2017) "Analysis of a Sequence of Events in Healthcare," *Journal of the Midwest Association for Information Systems (JMW AIS)*: Vol. 2017 : Iss. 2 , Article 2.

Available at: <http://aisel.aisnet.org/jmwais/vol2017/iss2/2>

This material is brought to you by the Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Journal of the Midwest Association for Information Systems (JMW AIS) by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Date: 07-31-2017

Analysis of a Sequence of Events in Healthcare

Pankush Kalgotra

Clark University, pkalgotra@clarku.edu

Ramesh Sharda

Oklahoma State University, ramesh.sharda@okstate.edu

William D. Paiva

Oklahoma State University, wpaiva@okstate.edu

Abstract

Healthcare industry generates streams of data in different problem domains. Analysis of such data requires stream analytics tools and techniques to generate useful insights. Stream analytics involve analysis of time variant events. The specific patterns in the events can indicate some imminent outcomes such as state of a heart, etc. Therefore, novel ways to find specific patterns in the events generated by multiple sources are required. A key requirement for applying any such method is data preparation and organization to enable such analysis. In this paper, we extend the CRISP-DM process to include data preparation approaches for sequence mining. We present progression analysis, an approach for converting multidimensional time variant streams of health records in a form to be able to detect useful sequential signals. To illustrate the process, we use patient health history stored in an Electronic Medical Record system (EMR) and present a healthcare application to compare progression of diseases over time between patients diagnosed with Tobacco Use Disorder (TUD) and non-tobacco users. Interestingly, many diseases follow the same path for TUD and non-TUD patients. Finally, the generalizability of the progression analysis is discussed.

Keywords: multidimensional time variant; streaming analytics; tobacco use disorder; sequence analysis, electronic medical records

Copyright © 2017 by Pankush Kalgotra, Ramesh Sharda, and William D. Paiva

1. Introduction

The proliferation of Big Data sources such as the web, sensor technologies, mobile technologies, etc. in healthcare has resulted into generation of large volume and variety of data. One of the sources of data in healthcare is Electronic Medical Record (EMR). An EMR contains patients' health information in electronic format (Hannan, 1996). EMR contains rich information and thus, many different Big Data analytics approaches can be applied to discover intelligence rooted in the data. Engendering intelligence through data analytics of EMR has the potential to help improve clinical practices, public health, hospital operations, etc., but many challenges arise in extracting this value. This paper focuses on one such promise and the associated data challenges.

Since EMR records time related clinical events of the patients, these provide opportunities to understand sequence of clinical events in terms of streams. A stream of data can be defined as a collection of time-ordered events. For example, if one disease leads to another disease, these diseases are recorded in the EMR ordered by time, and thus form a stream of data. EMR also contains lab procedures, lab reports and medication information recorded as streams.

In this paper, our focus is to analyze a multi-dimensional stream of data in EMR. A stream is multidimensional if there is a possibility of occurrence of multiple events at the same time. The multidimensional data streams are very common in practice. In an EMR, a data stream recording the patient information who develop multiple diseases at the same time and across multiple hospital visits is a multidimensional stream. We use the term multi-dimensional time-variant (MDTV) to term these data streams.

Storing, securing and analyzing such data streams has become one of the most critical issues for data scientists. To store and secure the huge volume and a variety of data streams generated with high velocity, efficient filtering and security measures are necessary (Chui, Löffler, & Roberts, 2010). Big Data tools and technologies are required to analyze massive data from variety of streams (Strohbach, Ziekow, Gazis, & Akiva, 2015). Since streams make a significant proportion of data in healthcare, it is important to develop methods to analyze these. To perform streaming analytics effectively and efficiently, a general process or methodology is necessary (Saltz, 2015).

The main purpose of this paper is to present and illustrate a method to organize multi-dimensional time-variant (MDTV) data streams in a form to perform sequence analysis (SA). Sequence analysis is one of the many methods to perform stream analytics and an efficient way to analyze time-ordered data (Agrawal & Srikant, 1995; Srikant & Agrawal, 1996). It is used to find sequential patterns in a transactional database. A transaction contains a set of items or events ordered by time. Sequence analysis has been widely used across the disciplines. In telecommunication, Asamoah, Sharda, Kalgotra, and Ott (2017) identified sequential activities before customer churn. The authors found that customers usually file complaints, using different communications channels such as call-center, online portal or physical store, before cancelling their service. A similar application of sequence mining has been presented in banking for finding patterns related to customer-loss by Chiang, Wang, Lee, and Lin (2003). Another common application of sequence analysis is in retail industry for discovering sequential patterns to understand the purchasing behavior as presented by Yun (2007). Mishra, Kumar, and Bhasker (2015) applied a similar approach for creating web recommendation systems. This method has also been used to find sequences in the career paths of information technology workforce by Joseph, Boh, Ang, and Slaughter (2012). There are also some applications of sequence analysis in healthcare for predicting patients who are at risk of developing a disease. For example, Batal, Valizadegan, Cooper, and Hauskrecht (2013) applied this approach to predict patients who are at risk of developing heparin induced thrombocytopenia. Similarly, sequence analysis can be a useful method to find patterns in EMR. Possible applications include discovering sequential activities in clinical events, progression of diseases, sequence of medications and many others. These applications require organization of EMR in a form to perform sequence analysis. In this paper, we describe a process to organize medical records for sequence analysis.

We adapt the traditional Cross-Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000) to analyze the time-stamped data streams. The adaption of CRISP-DM to analyze streams will extend its applicability to multiple other domains. CRISP-DM involves six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. All six phases of CRISP-DM are applicable to the streaming analytics. However, data preparation and modeling phases of the streaming data are different from the traditional static data mining due to its time-series nature. Researchers in the past have largely focused on proposing the models fitting into the data, which is important. But at the same time, novel approaches for preparing the time variant data are also required to organize data for modeling. As the data in raw form is difficult to analyze, it is crucial to convert the raw data in the form of a structured dataset to be able to perform any kind of analysis. Novel temporal methods based on the evolutionary algorithms are required as discussed by Gubbi, Buyya, Marusic, and Palaniswami (2013). The methods should be able to synchronize multidimensional data streams and generate meaningful patterns. We present a

process that will result into a dataset to find sequential patterns in the time-ordered transactional dataset. A similar algorithm has been explained by Eggho, Raïssi, Jay, and Napoli (2014), where the authors explained a process to find frequent patterns from a multidimensional sequential database.

We illustrate the proposed approach using data from an EMR to find how patients get diagnosed with diseases over time which may span multiple hospital visits. Since a patient can have multiple diseases simultaneously and may get diagnosed with multiple diseases across hospital visits, this problem of disease progression becomes a case of MDTV. First, we present a general method to convert the data streams from EMR in a form to be able to perform sequence analysis. Second, we illustrate the proposed method on a time-ordered dataset containing the patients diagnosed with Tobacco Use Disorder.

As discussed earlier, the primary purpose of our paper is to organize sequential data to perform data driven analytical modeling. The information based modeling is one of the topics of interest to Information Systems community as pointed out by Gupta (2017). The method presented in this paper is generalizable to other similar problems where information is required to be analyzed in sequences. Thus, its contribution to information systems literature is in proposing and illustrating an approach for such sequence analysis.

2. Method

To explain an EMR stream, we present a specific scenario in which a patient P gets diagnosed with diseases over multiple hospital visits. Moreover, multiple diseases may be diagnosed simultaneously during a hospital visit. A hypothetical scenario is presented in Figure 1 in which a patient visits hospital/s three times. He gets diagnosed with two hypothetical diseases A and B simultaneously during his first visit, followed by C and D during second visit, and finally E and F during the third visit. As more patients are added to the system in Figure 1, it gets complex and becomes difficult to understand the sequences of diseases.

Figure 1. A hypothetical patient's disease progressions across three hospital visits System

Patient	Visit no.	Diagnosis
P	1	A
P	1	B
P	2	C
P	2	D
P	3	E
P	3	F

Table 1. Cleaned and structured data example

2.1 MDTV Data Preparation Method for Sequence Analysis

To begin with, the dataset needs to be converted into semi-structured or structured form. The streaming data generated from the system in Figure 1 need to be cleaned and structured as presented in Table 1. In its extremely simplistic form, it includes the patient P, time of the event (hospital visit in this case) and diagnosis or event itself.

As our aim is to prepare the dataset for sequence mining, the streaming data is required to be converted into time-ordered unidimensional streams of events in the form of transactions. We adapt the concept of sessionizing to convert the multidimensional time-ordered streams into unidimensional transactions. The concept of a session is commonly used in web analytics to connect a series of user events online. A session is defined in multiple ways by the researchers in web analytics (Gayo-Avello, 2009; Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003). To list a few, a sequence of events

or queries are considered as one session if 1) these occur within a specific time interval, for example, one hour; 2) no more than time t passes between successive events. This time t is known as the “period of inactivity” or the “timeout threshold”; and 3) the collections of signals to complete a particular task are considered. In the MDTV case, any of the above approaches can be followed to sessionize the events. In our case, we adapt the third definition.

The steps followed to sessionize the medical records and find the progressions in the diagnoses across hospital visits are presented in Table 2. The process is illustrated with an example in Table 3a to Table 3d. To illustrate the process, we use a hypothetical dataset containing a patient (P) who visited the hospital three times (Visit Number.) as presented in Table 3a. On his first visit, the patient was diagnosed with two diseases (A and B); on the second visit, two new unique diseases were diagnosed (C and D); and on the third visit, E and F were marked. Our aim is to find diagnosis associations over hospital visits as in Figure 2.

To create a sessionized dataset appropriate for creating paths or sequences of diseases, we first divided the dataset into three sub-datasets, each having information about a visit as shown in Table 3b. The SQL left join of first visit dataset with the second visit dataset is performed, and its output is further left joined with the third visit dataset. In our hypothetical dataset, we have three visits. However, one can extend the analysis to any number of visits. The resultant table from multiple joins is presented in Table 3c, in which each combination of the diagnoses is labeled by a different session. Finally, the dataset is converted as in Table 3d by appending different visits. This dataset is now appropriate for sequence analysis.

The dataset prepared in Table 3d is the input for sequence analysis. The sessions in Table 3d are analogous to the transactions in the sequence mining method. The sequence analysis can be run to discover common sequence patterns. As the same disease can be repeated over time, an adjustment in the number of pair-wise combinations is made.

As the process involves joining of multiple large tables, it becomes difficult for the traditional data mining tools to analyze the data. Therefore, we used Teradata Aster, a Big Data platform. To run the analysis, the nPath function of Teradata Aster was used. This function creates paths with a session/transaction sorted by the time.

Step 1: For n patients ($i=1$ to n) with v visits and k diagnoses in each visit, the dataset contains

$P_i, V_{ij} (j=1 \text{ to } v), D_{ijk} (k=1 \text{ to } m)$

Step 2: A separate dataset for each visit is extracted; a total of v number of datasets are resulted:

$(P_i, V_{ij} (j=1), D_{ijk} (k=1 \text{ to } m)), (P_i, V_{ij} (j=2), D_{ijk} (k=1 \text{ to } m)) \dots (P_i, V_{ij} (j=v), D_{ijk} (k=1 \text{ to } m))$

Step 3: The dataset with $J=1$ is left joined with $J+1$. The resultant dataset contains

$(P_i, (V_{ij}, D_{ijk} (k=1 \text{ to } m))_{(j=1)}, (V_{ij}, D_{ijk} (k=1 \text{ to } m))_{(j=2)} \dots (V_{ij}, D_{ijk} (k=1 \text{ to } m))_{(j=v)}), (P_i, (V_{ij}, D_{ijk} (k=1 \text{ to } m))_{(j=1)}, D_{ijk} (k=1 \text{ to } m))_{(j=2)} \dots (P_i, (V_{ij}, D_{ijk} (k=1 \text{ to } m))_{(j=1)}, D_{ijk} (k=1 \text{ to } m))_{(j=v)})$. In each join, $D_{jk} \neq D_{(j-1)k}$, where $l=1$ to $(j-1)$.

Step 4: Each record is assigned a unique session. Total number of sessions per patient are

$S = \prod_{i=1}^v C_i$, where C is count of diagnoses in a visit i .

Step 5: A union of P_i, V_j, D_K and S_S are taken as

$(P_i, V_{ij} (j=1), D_{ijk} (k=1 \text{ to } m), S_{(1 \text{ to } s)}) \cup (P_i, V_{ij} (j=2), D_{ijk} (k=1 \text{ to } m), S_{(1 \text{ to } s)}) \cup \dots (P_i, V_{ij} (j=v), D_{ijk} (k=1 \text{ to } m), S_{(1 \text{ to } s)})$

Step 6: The Sequence Analysis is run to find the diagnosis-associations over time.

Step 7: The prevalence of a diagnosis-association is adjusted by calculating the number of patients following the specific disease path from the result of Step 3.

Table 2. Steps to prepare data for diagnoses progressions

Table. 3d

MDTV Sequence Analysis Process of EMR

3. Illustration

Our method can be applied to find sequences in the diagnoses in multi-dimensional EMR. To illustrate our method, we present the sequences of diseases over hospital visits diagnosed in the patients. We used medical records of patients from South-Central regions of United States (AR, LA, OK and TX) who were diagnosed with Tobacco Use Disorder. We obtained the dataset from the Cerner Corporation, a major Electronic Medical Records (EMR) provider. The database contains records of the visits of more than 50 million unique patients across US hospitals (1999-2014). In this paper, we focused on medical records of the patients diagnosed with the Tobacco Use Disorder (TUD).

3.1 Data Preparation

The step-by-step process followed to prepare and analyze the data is presented in Figure 3. The figure also includes number of records at each step. First of all, the information about the patients, hospitals, types of visits and diseases developed by the patients were combined for further analyses. There were more than 230 million records containing information about 50 million unique patients. Next, the emergency and inpatients diagnosed with the Tobacco Use Disorder in South-Central regions of US were extracted. At this point of time, we had 1.5 Million records containing medical records of 133,491 patients. As TUD is diagnosed at a certain point in the life of a patient, we considered his/her hospital visits from the first time occurrence of the TUD onwards. In the next step, we also made sure that a patient visited the hospital at least twice and TUD was diagnosed at least in two visits. These measures were taken to reduce the sample bias in the data as TUD may not be a prevalent diagnosis in a patient. If TUD is diagnosed at least twice, the disease sequences may be the indicative evidence of the TUD effect. After cleaning and preparation, the final dataset contained information about 25,330 patients. In the dataset, 56% of the patients were females; 75% were Caucasians, 19% African- American, 2.5 % Hispanics, 2.5% Native Americans and less than 0.2% were Asians. We applied our process described in the previous section on this dataset and organized it in the form of transactions to perform sequence analysis.

Figure 3. Data Preparation and Analysis Step

Moreover, to find the diagnosis-sequences in the Tobacco Use Disorder (TUD) patients, we compare the sequences in the patients who were never diagnosed with TUD (non- TUD). Similar to the dataset prepared for TUD patients, we extracted non-TUD patients from the pseudo-population who visited hospitals at least twice. The dataset contained information for 172,500 patients. It comprised of 62% females; 62% Caucasians, 28% African-American, 6.4 % Hispanics, 1.9% Native Americans and 0.6% Asians.

4. Results

In this section, first we present the sequences of diagnoses in tobacco users and then compare the associations across hospital visits in TUD patients with non-TUD patients. To illustrate the process, we restrict the sequence to the third visit in the hospital. So, maximum number of diseases in a sequence will be three.

4.1. Sequence of diagnoses in TUD patients

We present top twenty sequences in the diagnoses over three visits. The visualization of sequence of diagnoses is presented in Figure 4 and paths with percentage of patients following a specific sequence is listed in Table 4. Figure 4 is read left to right representing the three visits at each box. The width of an arc between two diagnoses represents its observed frequency of occurrence in our sample.

Figure 4 presents Hypertension (ICD9 – 401.9) as the most prevalent disease in TUD patients on the first hospital visit. It is followed by Diabetes mellitus, Hyperlipidemia, Coronary atherosclerosis or congestive heart failure on the second visit. Similarly, other sequences in Figure 4 can be interpreted.

Across the three visits, the most common sequence is Hypertension (ICD9-401.9), followed by Hyperlipidemia (ICD9-272.4) on the second visit and Diabetes mellitus (ICD9-250.00) on the third visit. (This can be observed in the first row of Table 4.) About 16% of the TUD patients who visited at least three times followed this specific path. The second most common sequence in Table 4 is Hyperlipidemia to Hypertension to Diabetes Mellitus. The third common sequence is Diabetes mellitus to Hyperlipidemia to Hypertension. These sequences show that Hyperlipidemia, Hypertension and Diabetes mellitus are highly related to each other across visits. Similarly, other sequences can be interpreted. As multiple diseases can be present at the same time and across the visits, one patient can exhibit multiple sequences and therefore, the sum of percentages is more than 100.

Figure 4. Sequence of diseases in TUD patients

S. No.	Sequence	Percent
1	[401.9, 272.4, 250.00]	16.27
2	[272.4, 401.9, 250.00]	15.49
3	[250.00, 272.4, 401.9]	15.47
4	[401.9, 250.00, 496]	14.65
5	[414.01, 401.9, 250.00]	14.63
6	[272.4, 401.9, 414.01]	14.24
7	[401.9, 250.00, 428.0]	14.24
8	[401.9, 272.4, 414.01]	14.12
9	[250.00, 401.9, 272.4]	13.96
10	[414.01, 401.9, 272.4]	13.95
11	[401.9, 250.00, 272.4]	13.86
12	[401.9, 414.01, 272.4]	13.75
13	[272.4, 250.00, 401.9]	13.73
14	[414.01, 272.4, 401.9]	13.64
15	[401.9, 414.01, 428.0]	13.64
16	[250.00, 401.9, 428.0]	13.35
17	[401.9, 414.01, 250.00]	13.31
18	[272.4, 414.01, 401.9]	13.03
19	[428.0, 401.9, 250.00]	13.03
20	[401.9, 272.4, 428.0]	12.94
ICD-9 codes		Disease
272.4	Hyperlipidemia	
401.9	Hypertension	
250.00	Diabetes mellitus	
414.01	Coronary atherosclerosis	
428.0	Congestive heart failure	
496	Chronic airway obstruction	

Table 4. Top 20 sequences of diseases over three visits

4.2. Comparison of diagnosis-associations in TUD patients with non-TUD patients

The sequence analysis in the previous section presented associations between diagnoses across three hospital visits. The results indicate that the same disease might appear across multiple hospital visits. Therefore, in this sub-section, we focus on the associations between diagnoses across hospital visits in TUD patients and compare with non-TUD patients. First, we present top 15 diagnoses associations in TUD patients across two visits and corresponding prevalence in non-TUD patients. Then, we describe the top 5 diagnoses during third visit and corresponding prevalence in non-TUD patients, given diagnosis associations across two visits.

4.2.1. Top 15 diagnosis-associations in TUD patients across two hospital visits and corresponding prevalence in non-TUD patients

Here, we describe fifteen most common pairs of diseases across first two hospital visits in TUD patients. In addition, we also compare these in non-TUD patients. The percentages of patients having a specific combination in both categories of patients are listed in Table 5. In addition, the same combination can be seen in the left box of Figure 5. Across first two hospital visits of TUD patients, the most common association is hypertension and diabetes mellitus. More than 7% TUD patients were diagnosed with hypertension and diabetes mellitus across first two visits. Interestingly, this association is common in non-TUD patients as well with 7% patients developing it. There are associations listed in Table

5 which are not significantly higher in tobacco users than non-TUD patients such as coronary atherosclerosis and diabetes mellitus, hypertension and congestive heart failure, diabetes mellitus and congestive heart failure and several others.

However, there are diagnosis-combinations that are highly prevalent in TUD patients. For instance, the pair of hypertension and chronic airway obstruction is more common in TUD patients than non-tobacco users. Similarly, pairs of diabetes mellitus and chronic airway obstruction, and coronary atherosclerosis and chronic airway obstruction are more prevalent in tobacco users. All other differences can be observed in Table 5.

Figure 5. Common disease-associations in the TUD patients over three hospital visits

4.2.2. Diagnosis-associations in TUD patients across three hospital visits and comparison with non-TUD patients

In our sample, about 51% tobacco users and 33% non-tobacco users visited hospitals at least three times. To find the associations across the three visits, we report the most prevalent diseases during third hospital visit given a pair is present in first two visits as reported in the previous section and Table 5. Given the top combinations across first two visits as described in Table 5, we report the top five diseases during the third visit. Each cell of Table 5 lists different diseases during third hospital visit with the two-visit associations. We report unique diagnoses observed during third visit and compare their prevalence in non-TUD patients.

Given the pair including hypertension and diabetes mellitus ($401.9 \leftrightarrow 250.00$) across first two visits, about 12% of patients who visited hospitals at least thrice get diagnosed as having a long-term use of medication (ICD-9 code: V58.69) and 10% patients reported long-term use of insulin (V58.67). Similarly, other diseases developed during the third visit in presence of specific combination can be interpreted from different sub-tables of Table 5.

Other unique diseases diagnosed in TUD patients during the third visit include old myocardial infarction, coronary atherosclerosis, shortness of breath, unspecified depressive disorder unspecified, chronic pain, peripheral vascular disease, anxiety disorder, unspecified atrial fibrillation, pneumonia, obstructive chronic bronchitis with exacerbation, personal history of venous thrombosis and embolism and percutaneous transluminal coronary angioplasty. These diseases can also be seen in the right box of Figure 5.

All numbers in the matrix are percentages. The numbers in green color represent the percentage of tobacco use disorder patients and the numbers in red-color represent the non-tobacco patients' percentage. Each cell in the matrix explains the diagnoses associations across two visits. The acronym "Preval" indicates the prevalence of the association. For example, the prevalence of 401.9 and 250.00 across two visits is 7.73% in tobacco users and 7.03% in non-tobacco users. The ICD9 codes in each cell of the matrix represent the top five diagnoses observed in the third visit given the two disease are observed in the first two visits.

*- *The difference between tobacco and non-tobacco patients is statistically significant ($p < 0.05$)

	250.00	272.4	496	414.01	530.81	428.0
401.9	Preval* 7.73 7.03 V58.69* 12.67 7.56 V58.67* 10.76 9.08 V45.81 9.90 9.31 786.50* 9.05 6.08 789.00* 8.10 5.70	Preval* 7.36 6.75 V45.81 15.16 13.91 V45.82* 11.98 7.90 412* 11.65 9.48 414.00 11.54 11.32 V12.51* 10.44 3.24	Preval* 6.09 2.09 491.21* 14.55 10.68 V45.81* 12.97 10.10 786.05* 11.15 8.93 311* 10.18 8.01 338.29* 10.06 2.93	Preval* 5.52 4.34 V45.82* 19.29 15.19 V45.81 19.15 18.20 412* 16.83 12.90 786.50* 13.27 8.07 414.00 12.59 12.56	Preval* 4.62 4.28 V58.69* 11.21 8.42 300.00* 10.60 6.19 786.50* 10.45 6.83 789.00* 10.45 7.04 V45.81 9.52 8.64	Preval* 3.58 4.11 V45.81 16.36 14.04 491.21* 15.43 4.28 786.05* 13.57 7.40 427.31* 12.64 26.15 414.00* 11.90 11.18
250.00		Preval* 3.31 3.65 V45.81 13.19 14.38 V58.69* 12.73 7.78 V45.82* 12.04 8.19 443.9* 11.81 5.87 412 11.11 10.02	Preval* 2.53 1.09 V45.81* 14.59 11.31 491.21* 13.79 10.71 786.05 11.67 9.87 443.39* 10.08 5.54 V45.82* 9.81 5.78	Preval 2.42 2.39 V45.82* 21.53 15.37 V45.81 20.96 19.64 412 15.01 13.32 414.00 14.73 12.98 443.9* 12.46 8.14		Preval* 1.95 2.71 491.21* 17.59 14.42 V45.81 15.96 15.02 V45.82* 14.01 8.22 786.05* 13.68 8.17 V58.67 13.03 11.07
414.01		Preval* 3.11 2.64 V45.82* 25.44 18.26 V45.81* 24.94 20.76 412* 20.91 14.85 414.00* 19.90 14.85 786.50* 14.86 8.36	Preval* 2.52 0.93 V45.82* 21.90 13.12 V45.81 21.33 17.96 412* 17.00 11.88 491.21* 14.12 8.98 786.50* 13.83 5.39			
428.0		Preval* 1.81 2.16 V45.81 21.13 20.33 412* 16.98 13.32 414.00 16.60 16.67 V45.82* 15.85 9.72 491.21* 13.96 4.36	Preval* 1.98 1.26 V45.81* 18.53 14.08 427.31* 17.25 26.34 491.21* 15.97 11.04 V45.82* 14.38 6.69 414.00* 14.06 10.33	Preval 1.92 2.09 V45.81 21.00 18.00 V45.82* 18.86 12.78 412* 18.86 13.08 414.00* 16.73 13.14 786.05* 16.01 9.40		

Table 5. Diagnosis-associations in tobacco and non-tobacco patients across three hospital visits

The prevalence of each pair of diseases is compared with the non-TUD patients in the red columns of Table 5. It is quite interesting to observe that some pairs are common in both TUD and non-TUD patients across first two visits but there are differences during the third visit. For instance, there is little difference in TUD and non-TUD patients with respect to the first reported combination across two visits in Table 5 (401.9 $\leftrightarrow$ 250.00) i.e. hypertension and diabetes mellitus. However, during third visit, long-term use of medication has more impact on tobacco patients than the non-tobacco patients. Similarly, other associations can be observed in different cells of Table 5 where there are large differences between two types of patient. These sequences need further research as past research has largely focused on the prevalent diseases in tobacco patients but not the disease-associations across hospital visits.

5. Discussion and Concluding Remarks

The above example is an illustration of the method of preparation presented in this paper. We showed the common sequences of different diseases developed by the tobacco use disorder patients over multiple visits. The sequences of diseases discovered can provide additional understanding about how different diseases develop sequentially in tobacco patients. The knowledge about sequence of diseases can help physicians provide additional understanding about the disease relationships based on time (Szentkirályi, Völzke, Hoffmann, Happe, & Berger, 2013; Kalgotha, Sharda, Molaka, & Kathuri, 2017) and take preemptive actions to prevent the future diseases. The progression analysis can be helpful to create an expert system for predicting future diseases.

We restricted our analysis to finding paths. However, it will be more interesting to see the outcome of a specific path. For instance, it is possible that some specific sequences of diseases can lead to death of a patient but not others. Classifying such sequences will be part of our future research. In addition, we only focused on sequences and associations across visits and not co-occurrences in the same visit. The co-occurrences provide static information. In contrast, our method considered time as an important factor. In this study, we primarily focused on the data preparation and analysis phases of CRISP-DM. Our future work will include evaluation of our method versus the existing methods and the deployment of our method in the real-world setting.

The application of progression analysis process is not limited to the healthcare problems. It is a generalizable process and can be applied to any situation in which multiple signals are generated over time. For example, finding patterns in the sequences of the signals generated by the IoT (Internet of Things) devices can be insightful. It is possible that a device in IoT behaves in a specific pattern just before its failure. An unexpected pattern of signals may forecast the future state of an IoT system. If such patterns are known and detected timely, preemptive maintenance actions can be taken. Since healthcare is a major opportunity for IoT systems, our process can help find sequence of events from different streams of health data. However, it has to be noted that the interval between visits in our healthcare example could be very large. In contrast, the interval between signals generated by the devices in the IoT environment is mostly very short. Hence, to apply progression analysis on fast generated signals, powerful Big Data tools are required. In our particular case, we used Teradata Aster, a Big Data Platform. In contrast with IoT environment, where all signals that are emitted are tracked by the system, only some of the patients' signals get noticed - only those that occurred during his/her visit.

Our study has few limitations. Analyzing the fuzzy events like diagnoses is a complex problem. Smoothening of such events by creating a time-boundary such as a hospital visit, is one way to manage such fuzzy events. Moreover, the actual time of a disease development is never known with certainty. A disease may have been developing in a patient, but it may not get diagnosed because the patient may not have been hospitalized, or even if hospitalized, clinicians could miss the diagnosis. Further, the patient may have gone to a different hospital where this specific EMR may not have been in use and thus the patient information is incomplete. We recognize this as a limitation of our study. In fact, this is a limitation of every study conducted using medical history of the patients. Also, the large dataset gives us an opportunity to address some of the data size issues. Further, by considering the entire hospital visit as a single time, we mitigate some concern of chronological order of diseases diagnosed during the same visit. Another limitation of the application of our approach on EMR is that we did not consider whether a disease disappeared over time. In addition, the time interval between the hospital visits was not considered. We also recognize that the method identified disease patterns over time. However, the co-existing diseases also form relationships and the techniques like network analysis or association mining can be applied to perform such analyses.

Notwithstanding the limitations, the sequence analysis can improve the decision making for physicians by presenting what other diseases can likely be diagnosed in future. Our contribution is in preparing the data to be able to perform modeling and analysis. Due to the growing streaming data in almost every discipline, generalizable methods such as one

presented in this paper can help enhance the power of data analytics.

Acknowledgment

This work was conducted with data from the Cerner Corporation's Health Facts database of electronic medical records provided by the Oklahoma State University Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Cerner Corporation.

* An earlier version of this paper was presented at 2016 IEEE International Conference on Big Data (Kalgotra & Sharda, 2016). In addition, parts of this paper have appeared in a chapter of book named "Guide to Big Data Application." (Kalgotra, et al. 2017).

References

- Asamoah, D., Sharda, R., Kalgotra, P. & Ott, M. (2017). Who Renews? Who Leaves? Identifying Customer Churn in a Telecom Company Using Big Data Techniques. *Journal of Information Systems Education*, forthcoming.
- Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering, 1995*(pp. 3-14).
- Batal, I., Valizadegan, H., Cooper, G. F., & Hauskrecht, M. (2013). A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(4), 63.
- Chiang, D. A., Wang, Y. F., Lee, S. L., & Lin, C. J. (2003). Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications*, 25(3), 293-302.
- Chui, M., Löffler, M., & Roberts, R. (2010). The internet of things. *McKinsey Quarterly*, 2(2010), 1-9.
- Egho, E., Raïssi, C., Jay, N., & Napoli, A. (2014, August). Mining heterogeneous multidimensional sequential patterns. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence* (pp. 279-284). IOS Press.
- Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179(12), 1822-1843.
- Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645-1660.
- Gupta, A. (2017). Editorial Thoughts: What and How ISR Publishes. *Information systems research*, 28(1), 1-4.
- Hannan, T. J. (1996). Electronic medical records. *Health Informatics: An overview*, 133.
- Joseph, D., Boh, W. F., Ang, S., & Slaughter, S. (2012). The career paths less (or more) traveled: a sequence analysis of IT career histories, mobility patterns, and career success. *MIS Quarterly*, 36(2), 427-452.
- Kalgotra, P., & Sharda, R. (2016). Progression Analysis of Signals: Extending CRISP-DM to Stream Analytics. In *2016 IEEE International Conference on Big Data* (pp. 2880-2885).
- Kalgotra, P., Sharda, R., Molaka, B., & Kathuri, S (2017). Time-Based Comorbidity in Patients Diagnosed with Tobacco Use Disorder. In *Guide to Big Data Applications* (pp 401-413). Springer International Publishing.
- Mishra, R., Kumar, P., & Bhasker, B. (2015). A web recommendation system considering sequential information. *Decision Support Systems*, 75, 1-10.
- Saltz, J. S. (2015, October). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In *2015 IEEE International Conference on Big Data* (pp. 2066-2071).
- Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13-22.
- Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Inform's journal on computing*, 15(2), 171-190.
- Srikant, R., & Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements*. Paper presented at the International Conference on Extending Database Technology.
- Strohbach, M., Ziekow, H., Gazis, V., & Akiva, N. (2015). Towards a big data analytics framework for IoT and smart city applications *Modeling and Processing for Next-Generation Big-Data Technologies* (pp. 257-282): Springer.

- Szentkirályi, A., Völzke, H., Hoffmann, W., Happe, S., & Berger, K. (2013). A time sequence analysis of the relationship between cardiovascular risk factors, vascular diseases and restless legs syndrome in the general population. *Journal of Sleep Research*, 22(4), 434-442.
- Yun, U. (2007). Analyzing sequential patterns in retail databases. *Journal of Computer Science and Technology*, 22(2), 287-296.

Author Biographies

Pankush Kalgotra is a doctoral student at Oklahoma State University in Management Science and Information Systems department. His primary research interests include data analytics, network science, dark side of IT and neuroimaging in Information Systems. He is proficient in using and teaching Big Data tools such as Hadoop and Teradata Aster.

Ramesh Sharda, Ph.D. is the Vice Dean of the Watson Graduate School of Management, Watson/ConocoPhillips Chair and a Regents Professor of Management Science and Information Systems in the Spears School of Business at the Oklahoma State University. He also serves as the Executive Director of the PhD in Business for Executives Program. He has co-authored two textbooks (Business Intelligence and Analytics: Systems for Decision Support, 10th edition, Prentice Hall and Business Intelligence: A Managerial Perspective on Analytics, 3rd Edition, Prentice Hall). His research has been published in major journals in management science and information systems including Management Science, Operations Research, Information Systems Research, Decision Support Systems, Interfaces, INFORMS Journal on Computing, and many others. He is a member of the editorial boards of journals such as the Decision Support Systems and Information Systems Frontiers. He is currently serving as the Executive Director of Teradata University Network and received the 2013 INFORMS HG Computing Society Lifetime Service Award.

William Paiva, Ph.D. is the executive director of Oklahoma State University's Center for Health Systems Innovation (CHSI). William is a health and biotech venture capitalist who served as a board of director for the CHSI since its establishment in 2012. He is an Oklahoma native who received a Ph.D. in Molecular Biology from the University of Oklahoma and an MBA from Tuck School of Business at Dartmouth College. William brings a broad range of knowledge and expertise from his management consulting experience with Pittiglio, Rabin, Todd, & McGrath, providing business strategy and process consulting services to large pharmaceutical companies, as well as from his tenure as an associate in J.P. Morgan's Healthcare Investment Banking group in New York City. For nearly 20 years, William has invested in cutting edge health care companies through Sevin Rosen Funds, Chisholm Private Capital and the Oklahoma Life Science Fund. In addition, he's also taught venture capital courses and entrepreneurship classes at OSU's Spears School of Business, Tulsa University, and the University of Oklahoma.

Page intentionally left blank