Jyotsna Krishna Sastrula. Analysis and Visualization Methods for Data-Driven Longitudinal Patient Summary. A Master's paper for the M.S. in I.S. degree. May, 2016. 35 pages. Advisor: David Gotz

Digitization of health records has opened avenues for intensive research in the fields of health informatics. Power of machine learning, statistical analysis and visual analytics could be utilized to make optimal use of this information. The proposed project is to develop an interactive visualization tool that summarizes a patient's medical history, highlighting all his/her important events based on the knowledge of similar patients. Given a set of patients with common conditions, statistical analysis can be used to develop models that prioritize features based on associations between features and condition-specific outcome measures.

This manuscript in particular describes the model developed to prioritize a patient's events from his medical history. The model is trained with the population of patients and their events. Their correlations with the outcome variable are calculated to identify the important events in a specific cohort. This correlation score can be used to prioritize the events associated with an individual patient. This model is one of the models that will be used to summarize an individual patient's medical data via interactive visualization methods.

Headings:

Electronic Health Records

Statistical Analysis

Visual Analytics

Predictive Models

ANALYSIS AND VISUALIZATION METHODS FOR DATA-DRIVEN LONGITUDINAL PATIENT SUMMARY

by
Jyotsna Krishna Sastrula

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

May, 2016

Approved by:

_____

David Gotz

Table of Contents

1. Introduction:

When a patient visits a doctor, usually the appointment is timed to be for 20 minutes [1]. It is critical for the doctor to be informed of any prior medical conditions that the patient had which could be relevant to the current complaint he/she has. Luckily, doctors have access to the medical history of a patient. But in the given timeframe of 20 minutes, it is practically impossible for a doctor to go through the history of the patient and identify the important medical conditions he/she has had. It is unfortunate that all the required data is available but is not in a usable format leading to some bad decisions and fatal results in some scenarios. In the current scenario, to avoid misunderstandings and incorrect diagnoses, a doctor must ask the patient right questions before coming to a decision. This cannot be done every single time and that one time, when the doctor does not ask correct questions, can turn fatal. Also, there can be cases where the doctor asks right questions but the patient does not know/remember to tell the right answers. Few times, there can be similar patients who were already diagnosed leading to either good or bad results. Having this information handy assists a doctor in deciding on correct diagnosis.

All the above mentioned problems can be resolved to some extent with the power of analytics and proper interactive visualization. We are developing a tool that addresses the above mentioned problems by exploiting the power of analytics and visualizations. The tool populates a personalized, context-appropriate visual summary of a given patient's medical history. Doctors can utilize this tool to take informed treatment

decisions for a given patient's current condition. We will use several statistical tools to develop a model which analyzes the history of a given patient and identifies any earlier context-appropriate conditions he/she might have had. Then this data is summarized visually using various visualization tools.

2. Literature Review:

2.1.Clinical Behavior:

My Literature review plan is to include literature on the consultation lengths of the doctors and behavior studies on doctor-patient communication during a consultation. A review of the related work on mining data in Electronic Health Records and various visualization methods is also presented in this section.

Wilson [1] examined historical and international comparisons of consultation length. He reported that the mean lengths are: 10 minutes in the United States of America, 12 minutes in New Zealand, 15 minutes in Canada and 21 minutes in Sweden. The determinants of the consultation length included both the variation between doctors and variation between patients. Explanatory variables for 'variation between doctors' were age, sex, training and attitudes of the doctor, and the practice list size. He noticed that older doctors have longer consultations and also that women doctors have longer consultations than men. The observation for the 'variations in patients' was that consultations about new patient problems were longer than those for known problems (mean of 5.8 minutes compared with 5.2 minutes). This review has studied the evidences that longer appointments prescribed less(51.2% as opposed to 62.6% compared to the lesser appointment practices) and patient initiated revisits over the subsequent four weeks is also less(7.2% versus 12.9%). The review concluded by saying that in order to achieve

longer consultations general practitioners would have to reduce their list size, decrease their patients' consultation rate with the doctor (by increasing delegation), or work longer hours. It suggests that doctors should be enables to consult at a pace that suits them within an appointment system.

Though the above study seems a little old, the analysis done by Mechanic et al [2] shows proves that there is no significant change in the duration of a patients office visits with the physician for a decade (1989 – 1998). Thus we can safely assume that the consultation lengths have remained the same over time. They used data from two nationally representative sources - National Ambulatory Medical Care Survey (NAMCS) of the National Center for Health Statistics and the American Medical Association's Socioeconomic Monitoring System (SMS). In 1998, the average duration of office visits was 16.3 minutes as per NAMCS data and was 20.4 as per the SMS data. According to both the sources of data, the average duration has increased by 1-2 minutes between 1989 and 1998.

In the available consultation time, it is very common that a patient does not voice all his expectations in a visit and this might lead to undesirable consequences [4]. These consequences can turn out to be fatal in the cases where a patient does not inform the doctor of any previous medical conditions that might be related to the current condition. This leads to a misunderstanding leading to undesirable diagnosis, non-adherence to treatment etc. Britten et al [3] have studied the misunderstandings in prescription and identified 14 categories of misunderstandings related to patient information unknown to the doctor, doctor information unknown to the patient, conflicting information, disagreement about attribution of side effects, failure of communication about doctor's

decision, and relationship factors. They interviewed 35 patients before and after the consultation. They also recorded the consultation. They analyzed this data and came up with the above mentioned categorization. They said that though the doctors are tempted to assume that they know their patients well, this not the true in most cases. Apart from listening, doctors have to ask right question in order to avoid undesired prescriptions and thus avoid the resulting adverse effects.

Our project focuses on minimizing the misunderstandings related to 'patient information unknown'. It learns, from similar patients' information, what events are relevant and important that are associated with an outcome. This knowledge is then used to summarize a given patient's medical history, highlighting all his relevant events. This helps a physician to understand the health information of a particular patient. He will not have to "assume" things about a patient and he can ask right questions.

In another study, the authors reviewed literature on medical errors and preventable adverse events in primary care to classify medical errors. They searched MEDLINE and the Cochrane Library from 1965 through March 2001 to identify relevant literature. They derived a classification system with two categories - Classification of preventable adverse events in primary care & Classification of process errors in primary care. The first classification is comprised of –Diagnosis (Related to symptoms, Related to prevention), Treatment (Drug, Non-drug) and Preventive services (Inappropriate, Delayed, Omitted, Procedural complication). The classification is comprised of - Clinician factors (Clinical judgment, Procedural skills error), Communication factors (Clinician-patient, Clinician-clinician or health care system personnel), Administration factors (Clinician, Pharmacy, Ancillary providers, Office setting) and Blunt end factors

(Personal and family issues of clinicians and staff, Insurance company regulations, Government regulations, Funding and employers, Physical size and location of practice, General health care system). It is clear from this classification system that the first classification tells 'what went wrong' and the second classification tells 'why something went wrong'. Upon closer observation, we can conclude that both kinds of the errors can result from physician being not completely informed of the patients' health history. Hence we believe that having a tool at a physician's hand which informs him of a patient's health history and also highlights any important (good and bad) events that resulted due to a certain diagnosis can help him take more informed decisions and thus avoid adverse events that stem from bad diagnosis or wrong judgement.

2.2.Mining Data in Electronic Health Records (EHRs):

With the availability of electronic health records, there is a lot of evidence present for a certain condition and its various treatment effects. There are a lot of studies that developed predictive models utilizing EHR data that identify various patterns and predict a certain outcome. For example, Kurosaki et al [15] developed a model that identifies patients at high risk of developing Hepatocellular carcinoma (HCC) within 5 years. They collected various test data of 1003 chronic hepatitis C patients for a period of 5 years. Since theirs was a prediction problem, they developed a decision tree model which searched their analytical database for the factor that most effectively predicted HCC development and for its cutoff value. Another study [18] builds a predictive model that uses the populations of patient data to predict the mortality of a given patient in ICU on their 5^{th} day of stay. As opposed to the earlier study, the problem here does not consider a

prolonged information of the patient in question, but concentrates on similar population of patients to make predictions.

While the above studies concentrated on finding solutions for a specific problem using data mining and predictive analysis, there are several studies which developed tools that could be used for a wide variety of problems. An example of one such tool is *iHealth Explorer* tool [17]. It allows users to choose from a collection of datasets and their analysis need. The tool then applies various analyses and provides insights on the data for the specific need. PatternFinder [16] is another tool that allows users to search for patterns in EHRs. It allows makes temporal querying easy for clinicians.

When it comes to temporal data, a lot of complexities come into picture as the nature of the data in itself is complex and is computationally highly demanding. It difficult to mine temporal data and several studies have proposed few solutions. As our problem involves mining temporal longitudinal data of patients, literature in this field is relevant. One such [19] study presents an algorithm – KarmaLego which is fast and enumerates Time Interval Related Patterns TIRP from temporal longitudinal data. In [20], the researchers present a method - One-Sided Convolutional Nonnegative Matrix Factorization (OSC-NMF), to extract temporal patterns from longitudinal data. Their framework mines common as well as individual temporal patterns from heterogeneous events.

There still is a lot of research going-on on mining populations of patient data and extracting informational patterns that help in clinical decisions. But considering the fact that mining a particular dataset biases the results towards the dominating ethnic group the dataset is comprised of, but is not generalizable, there are studies that are advocating

personalized medicine. This is emphasized in [21], where the author says that the observed patterns in a population cannot be applied to every individual as their demographics vary largely like the ethnicity, age and gender. They suggest N-of-1 trials which means, collecting enough data for a single patient for a long time and identifying the patient as responder or non-responder to a treatment. Aggregating results of many such N-of-1 trails can yield information on subsets of population. The proposed project in this paper can to some extent utilize this concept as it summarizes the response a single patient to a particular treatment. The data of such information of a population of patients can be considered in identifying a subset of population belonging to a particular cohort.

2.3.Visual Analytics:

Mining of electronic health records (EHRs) has the potential for establishing new patient-stratification principles and for revealing unknown disease correlations [6]. In this article authors elaborate the importance and usefulness of mining Electronic Health Records. It enhances Clinical Decision making and enables informed decision making. They emphasize on research need in temporal data analysis mainly on mining longitudinal patient data and establishing patterns which could be used for predictive purposes. Similar view is also expressed by the authors of "The inevitable application of big data to health care."[7]. They list out the advantages of applying big data to health care and while doing so, they state that though the accessibility to latest clinical studies provides evidence guiding clinical practice but the sheer volume of the information makes it difficult to transform this information to knowledge. This is where our tool comes in as a tool which analyzes similar patients and highlights the medical conditions that are critical to particular patient by analyzing his longitudinal data.

Visual analytics was chosen because, it integrates the human visual cognitive abilities with the power of statistical analysis and interpretation of knowledge. It is believed that visual representation of the information reduces complex cognitive work needed to perform certain tasks [8]. They presented the visual analytics mantra “analyze first - show the important - zoom, filter and analyze further - details on demand". This is exactly what our plan for the project is. We will be analyzing the data using statistical tools, visualize a patients historical medical data, show important events associated which will be calculated using other similar patients’ data, these important events can be zoomed and further analyzed to get more details before a doctor decides on a diagnosis to be administered to the patient. In another literature [9], that narrates the numerous applications of analytics and clinical informatics in health care, visual analytics has been given the top priority. According to them – “There are three main benefits to the visual analytics approach versus the traditional method of querying databases. First, the user can explore the data in a self-service fashion, as opposed to writing database queries by hand. Second, complex ideas can be communicated with clarity, precision and efficiency in visual graphs, rather than the tabular data output from a traditional database query. Third, visual analytics can display large volumes of filtered data in near-real time, which is a more onerous task when using traditional database queries”. They studied literature which exploited the power of visual analytics to communicate complex data effectively and support clinical decision making.

Medical data such as that contained within a patient’s medical history is highly temporal in nature. For this reason, research related to the visual analysis of time-varying data is highly relevant, and it is a well-studied subtopic within the visual analytics

literature. In [10], the authors present a systematic view of methods visually analyzing temporal data. According to them there are three questions that correspond to the temporal visualization categorization criteria: time, data and representation –

(1) What are the characteristics of the time axis? – can be time points or time intervals which are either linear or cyclic or branching.
(2) What is analyzed? – The data that has to be tied to the time axis. This can be abstract or spatial. Or it can be univariate or multivariate.
(3) How is it represented? The representation can be static or dynamic with either 2-dimensional or 3 dimensional presentation.

We believe that this systematic view will help us brainstorm on how to present our data on a time axis. Our data is a mixture of time points and time intervals and is n-dimensional in nature. We have to abstract this n-dimensional data to a 2D visualization.

One of the seminal works which focuses on visualizing personal history is [11]. The authors here propose a technique to visualize personal history – LifeLines. It allows to visualize multiple facets of a person's life to be visualized like medical, financial, legal, etc. They present each event over time as a horizontal line whose thickness and color are used to indicate the severity of the offense and the depth of penetration in the system (when visualizing legal data of a person). Their application to medical records is what we are concerned about. They show events in horizontal lines with appropriate labels. As the severity of the condition reduces due to diagnosis, the width of these lines is also reduced. They used different icons to show different kinds of information. They provide a complete visualization environment offering overview, zooming, filtering and details on demand. Paper [12] builds on this work and propose space efficient

visualization which is abstracted at various levels which can be uncovered as required. Their technique connects *overview+detail*, *pan+zoom*, and *focus+context* features to one powerful time-browser. They suggest an abstraction of information in color and height coded horizontal lines by combining LifeLines and *Graphical Summary of Patient Status* techniques. Data can be resized vertically through the different visualization techniques and abstraction levels of the data, adding details step by step. Their time visualization is spread over three connected time lines. The first (bottom) one provides a fixed overview of the underlying data and its full temporal range. Selecting a sub-range in the first timeline defines the temporal bounds for the second (middle) and third (top) timeline. By interacting with that sub-range you can easily *pan+zoom* in time.

All the relevant literature describes only the visualization techniques but none talk about how the data behind the visualizations are stored and accessed. This project is to first develop a model which ranks medical events that are associated with a particular cohort and highlight these on a particular person's history summarization. The model development and ranking need statistical analysis whose output is connected to the time axis on the summary.

3. Methods:

The goal of this project is to develop a model that can highlight the important events associated with a given patient. The model will be trained using the data related to similar patients and their events. The model will calculate the correlation score of each of the events with the outcome variable. These correlation scores are used to rank the events associated with the given set of patients. Once the events are ranked, these ranks will be used to highlight the events associated with a given patient who is similar to the set of

patients on which the model is trained. The following sections describe the data used to train the model, the features that are extracted out of the data and the algorithm used to build the model.

3.1.Data:

To build an initial model, data from the MIMIC II Clinical Database [23] has been used. This database contains tens of thousands of Intensive Care Unit (ICU) patient data. The data were collected between 2001 and 2008. It has Patient Demographic data, death dates, ICD-9 codes etc. all the dates are surrogate dates due to privacy issues.

Data has been organized into tables. This data is comprised of all the patients' events recorded like Admitted into the hospital, discharged from the hospital, any procedures performed in them etc. For this study, the events and procedures performed on the patients is of interest rather than any demographic information about them. All the events are maintained in a dictionary table in the database which also has the ICD-9 codes mapped to each event. Each patient's events are recorded in a transaction table with the event_ids and the timestamp at which this event has occurred. Following ER diagram explains the table structure used to store the patients data. Knowing this information helps in extracting the features needed for the model's training.

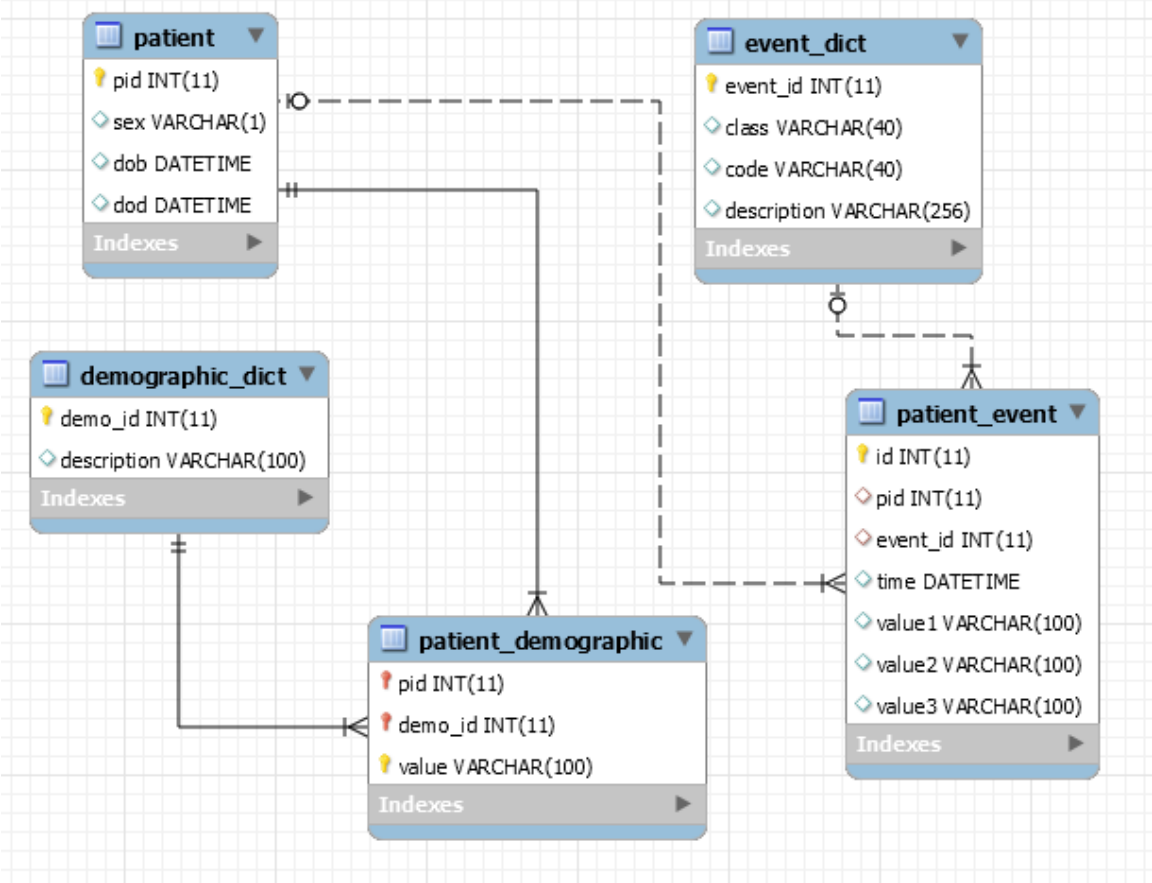

Figure 1: ER Diagram of the tables used to store the patient information used for the project

All the patients with their basic information like sex date of birth and date of death are stored in the patient table. 'dod' column is null for the patients that are not dead. Any other demographic data like their ethnicity, Marital Status, Religion, etc. are stored in the patient_demographic table. This table also has 'EXPIRED IN HOSPITAL FLAG' information for each of the patients. This information is used as outcome variable in this study. All the event_ids are treated as features for this study. The event_dict table has all the event types associated with all the patients in this dataset is. There are 3341 distinct event types recorded in the dataset. Patient_event table is a transaction table that

has an entry for each of the events that occurred to every patient. This is the most important table which is used to build our feature matrix.

3.2.Feature Description:

The aim of this study is to develop a model which will take a patient_id as an input and outputs a ranked list of events from his history. Already available data about other similar patients and their outcomes as a result of a treatment is indicative of what events in a particular patient's history are important. For this study, a patient's mortality is considered as the outcome variable. The feature matrix is the numerical matrix with event_ids as the rows and patient_id as columns. The values being the number of times a patient had a particular event. For example, if the element at row M and column N is 2, that means that the patient N had the event M 2 times. There are 32535 distinct patients and 3341 distinct events available. Hence our feature matrix will be a 3341x32535 matrix. The outcome vector is a matrix one row for each patient and the binary value of either 1 or 0 implying if the patient expired in the hospital or not. As there are 32535 distinct patients, our feature vector is a 32535x1 matrix. Once the feature matrix is built, the correlation of each of the events with the outcome variable will be calculated. This correlation score is used to rank the event types.

As can be seen, the feature matrix is very large. As a result, the time taken to process all this data was approximately 20 hours. Therefore, a feature reduction step was required to allow timely execution. Feature reduction is also important to avoid over fitting the model. To identify the relevant features, the patient_event table has been analyzed. There are a total of 3341 distinct event types recorded in the event dictionary table but only 1774 event types have been reported by patients. All the events that have

event_id between 60001 and 90031 were never reported and belonged to the classes ‘HFCA_DRG’ and ‘MICROBIOLOGY’. As these events have no useful information, these were removed from the feature matrix. This resulted in the feature matrix of size 1774x32535. All the features fall into two classes: ‘STATUS’ and ‘PROCEDURE’.

	class	count(distinct pid)
▶	HFCA_DRG	0
	MICROBIOLOGY	0
	PROCEDURE	29126
	STATUS	32074

Figure 2: Distribution of Patients across various Classes of Events

	class	count(distinct event_id)
▶	HFCA_DRG	1018
	MICROBIOLOGY	436
	PROCEDURE	1885
	STATUS	2

Figure 3: Distribution of Events across various classes

Similarly, there are 461 patients with no events recorded at all. These patients are also removed from the feature matrix and the outcome vector. This results in the feature matrix’s size to be 1774 x 32074 and that of the outcome vector to be 32074x1.

The distribution of the data with respect to the outcome variable is unbalanced. The number of patients data with outcome variable value = N is extremely higher than the one’s with the value = Y.

	value	count(distinct pe.pid)
▶	N	28629
	Y	3445

Figure 4: Distribution of data with respect to outcome variable

This means that there are not enough examples for the outcome with a value Y. Since we are using correlation score between the features as the main function of the model, this should not be an issue. The model does not depend on the balance of the dataset.

3.3.Model:

The model takes a patient ID as an input and the output is a list of his/her events ranked as per their correlation with the outcome variable. Mortality of a patient is considered as the outcome variable in this project. The outcome variable can vary and a different model will be needed for each of the outcome variables. Python's numpy, scipy and pandas packages are used in the project. Numpy is used to build various complex and huge arrays required for the computation, pandas is used for transformation of the huge data as per the requirements and scipy is used for statistical analysis. Feature matrix is built by fetching records from the MySQL database. Python's pymysql package is used for the database connectivity. The records fetched are returned as tuples which then are converted into a matrix with event_ids as rows and patient_ids as columns and the corresponding value to be the number of times a patient had a particular event. Similar data transformation is applied for the outcome vector as well. Once the feature matrix and the outcome vector are built, scipy's stats module is used to calculate the correlation of each feature with the outcome. Each event's values corresponding to all the patients and the outcome of all the patients are fed as the observations to the pearsonr method and the correlation for each event is extracted and stored in a matrix along with the event_id. Once the total correlation matrix is built, the events are ranked based on their correlation scores. This is a one-time calculation. Once all the events are ranked, these can be used to

rank the events of a given patient. All the events in the medical history of a given patient are ranked taking the ranks of those events from the correlation matrix.

Scipy.stats.rankdata function is used to rank the events based on their correlation score. Event with least correlation is ranked as 1 and the one with the highest correlation is ranked the highest. Dense ranking method is used which means, in case of ties, the rank of the next highest element is assigned the rank immediately after those assigned to the tied elements. Each of the tied elements will have the minimum of the ranks that would have been assigned to all the tied values. A positive correlation score means that the feature is likely to result in the death of a patient in the hospital and a negative correlation means that the feature results in a good outcome.

This list can be used in the visualizations to highlight a given patient's events by reading the rank of the events. The output has the ICD-9 codes, which are the standards used in medical field, mapped with the event_ids.

3.4.Algorithm:

Given a set of patients with a common condition, statistical analysis techniques are used to develop models of feature priority based on associations between the features and condition-specific outcome measures. This model is then used to prioritize a given patient's conditions.

In general, the model fetches the required data from the database, transforms it into matrices to be used for the statistical analysis and then calculates the correlation between the features and the outcome variable. It then ranks these features based on their correlation scores. Data cleansing is taken care of in the database itself while fetching the

required data. Transformation involves moving data into appropriate data structures to be usable for the statistical analysis. Statistical analysis mainly involves calculating the correlation scores of each of the features with the outcome vector and then ranking these features based on their correlation scores. Once these as features are the event_ids which are not meaningful in the medical world, these event_ids are mapped back to their ICD-9 codes from the database. Once the model is ready it can be used to prioritize events of a specific patient. The general algorithm used to develop the model can be summarized as follows:

Step 1: Load and Transform the Data

- Load the cleanse data from the database
- Initialize a matrix with events as rows and patients as columns
- Transform the data fetched and fill the matrix initialized with the number of times a patient had a particular event
- Similarly fetch cleansed data from the database for Outcome Vector
- Initialize a matrix with patients as rows
- Fill the matrix by reading the patents demographic data related to their death in the hospital. Wherever the value is ‘Y’, it is denoted as 1 in the outcome vector and 0 otherwise

Step 2: Calculate the Correlations

- Initialize a correlation matrix with the number of events as size and 1 column

- For each event, calculate the correlation with the outcome variable. The number of observations will be the number of patients
- Fill in the correlation for each event in the correlation matrix along with the event_id

Step 3: Map the Events to ICD-9 Codes

- Fetch the event_id and ICD-9 code mapping from the database into a list
- Initialize an empty list
- Loop through the correlation matrix for each event_id. Get the corresponding ICD-9 code from the fetched list. Fill the initialized empty list with [event_id, associated ICD-9 code and the rank of this event] as one entry.

This algorithm works well with huge datasets. With the data used for this study, which involved 1774 features and 32074 observations, the algorithm took around 5 minutes on an average to complete the execution and learn the model. This includes the time to connect to the database over network, fetch the records, transform them, statistical analysis and then mapping the results to appropriate medical codes.

4. Results and Discussion:

The output of this model is the list of events with their correlation scores and ranks. This list is prepared once and then used to rank the conditions of a given patient each time. It is important to note here that the results obtained in this study are not from the intended longitudinal patient data. The dataset used is similar to the intended dataset but is collected for a different purpose and hence does not have the historical details of each patient. As a result, the events recorded in the data are not diseases or conditions but

are procedures performed on each patient. These are not the causal conditions for a patient's mortality. But one can say, if a patient reaches to point where these procedures have to be performed, then the chances of mortality are high. These are the procedures that are highly correlated with the mortality of a patient. The model is a list of 1774 distinct events with their ICD-9 codes and ranks. Using this model, a specific patient's events can also be ranked. An event with lower rank is of lower priority and an event with higher rank is of higher priority.

The maximum correlation score that achieved is 0.2577 and the most negative correlation score achieved is 0.1602. Following histogram shows the number of events per correlation score:

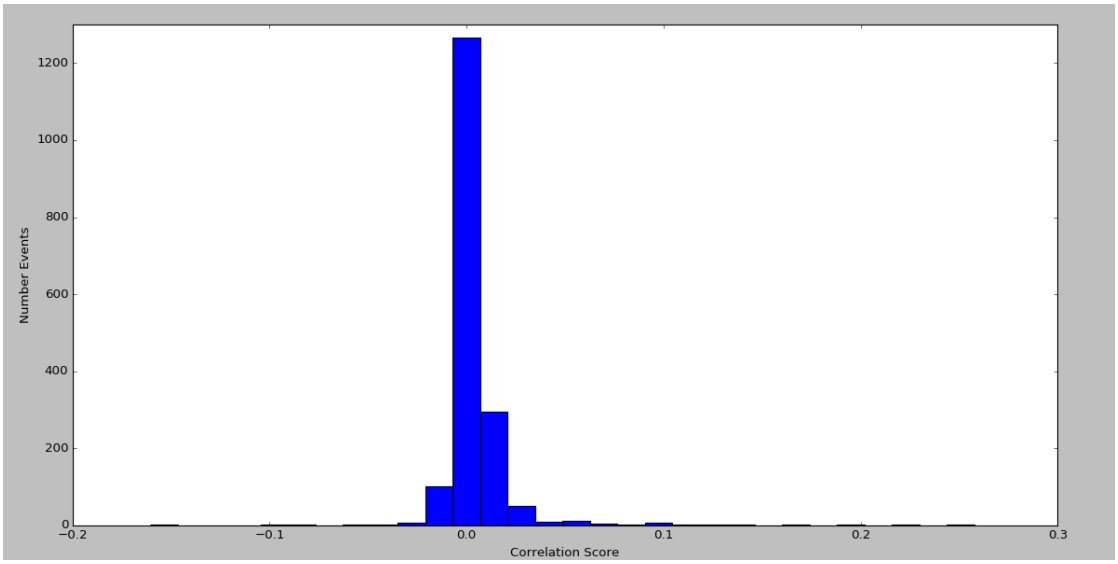

Figure 5: Histogram showing number events per correlation score

Most of the events have a correlation score of zero. Out of the 1774 distinct event types, there are around 1250 events with a correlation score close to zero. This is because most of the events are mostly procedures and are not associated with the differences in outcome variable.

Following is the table of highest ranked 10 events with their ICD-9 codes and their correlation scores:

event_id	ICD-9 Code (DESCRIPTION)	Correlation Score
101749	9604 (INSERT ENDOTRACHEAL TUBE)	0.2577
101783	9672 (CONTINUOUS INVASIVE MECH)	0.2231
101782	9671 (CONTINUOUS INVASIVE MECH)	0.1887
101866	9960 (CARDIOPULM RESUSCITA NOS)	0.1741
100574	3893 (VENOUS CATHETER NEC)	0.1738
100572	3891 (ARTERIAL CATHETERIZATION)	0.1721
100009	0017 (INFUSION OF VASOPRESSOR)	0.1362
101837	9907 (SERUM TRANSFUSION NEC)	0.1232
101835	9905 (PLATELET TRANSFUSION)	0.1150
101780	966 (EXT INFUS CONC NUTRITION)	0.1135

Table 1: Highest ranked 10 events

The maximum correlation score achieved is 0.2577 and it is for the procedure 'INSERT ENDOTRACHEAL TUBE'. This is a procedure performed often when patients are critically ill and cannot maintain adequate respiratory function to meet their needs. The endotracheal tube facilitates the use of a mechanical ventilator in these critical situations. Any condition of a patient which leads to adoption of this procedure can be viewed as highly correlated with mortality. The next highly correlated procedure is

CONTINUOUS INVASIVE MECH. This is a procedure performed when a patient is unable to breathe. Invasive here indicates that an endotracheal tube is inserted. When observed all the procedures listed in the above table are all risky procedures and are usually adopted while treating life threatening conditions. Hence, the high correlation with mortality.

Following is the table of least ranked 10 events with their ICD-9 codes and their correlation scores:

event_id	ICD-9 Code (DESCRIPTION)	Correlation Score
101645	8853 (LT HEART ANGIOCARDIOGRAM)	-0.031415707099126874
100478	3722 (LEFT HEART CARDIAC CATH)	-0.03551867875741903
100423	3521 (REPLACE AORT VALV-TISSUE)	-0.036820095002895306
100463	3613 ((AORTO)CORONARY BYPASS T)	-0.049590568718116505
100462	3612 ((AORTO)CORONARY BYPASS T)	-0.05166983988873123
100465	3615 (1 INT MAM-COR ART BYPASS)	-0.08207961247495588
101881	9983 (OTHER PHOTOTHERAPY)	-0.08747859784604059
100606	3961 (EXTRACORPOREAL CIRCULAT)	-0.09070583933084606
101078	640 (CIRCUMCISION)	-0.09149027327561951
101863	9955 (VACCINATION NEC)	-0.16023815525102855

Table 2: Least ranked 10 events

It can be seen that these events are negatively correlated, meaning that these procedures are mostly associated with patients that who have not expired in the hospital. The procedure with the most negative correlation score is VACCINATION NEC. This procedure mean a vaccination has been given to the patient. As said above, because the dataset used is not one intended for the study, the events recorded are not exactly related to COPD, HF or DIAB. Hence the results do not immediately seem intuitive. However, on deeper observation, the results seem meaningful. In this case, vaccines are administered to infants only as a preventative measure and hence any procedure related to vaccinations is negatively correlated to mortality. These records could be of infants who are actually healthy. However, this could not be verified from the dataset as the data is masked as per IRB specifications to make it unidentifiable.

An interesting procedure in this result is EXTRACORPOREAL CIRCULAT, which is diversion of blood flow through a circuit located outside the body but continuous with the bodily circulation. This is adopted while performing an open heart surgery. This sounds like a risky procedure but actually has a negative correlation with mortality. We are hoping for some interesting results like this after we apply this model to the dataset with historical patient data. This kind of results help doctors with their treatment decisions. Having information of treatments that resulted in positive outcomes with similar other patients makes it easier for the doctor to suggest similar treatments to the current patient and can actually get positive results too.

Consider patient with ID = 12, who is dead, as an example. When we pass this patient ID as an input to the model developed, a list of all his event_ids along with their

ICD-9 codes and the ranks of these events is expected as an output. Patient 12 has the following events recorded.

pid	event_id	code
12	100574	3893
12	100881	5137
12	100905	5212
12	100939	5351
12	100952	5412
12	100962	5459
12	101749	9604
12	101782	9671
12	101843	9915
12	101866	9960
12	1	001
12	2	002

Figure 6: Events and their ICD-9 codes recorded for Patient 12

These events were ranked by the model as the follows:

[[100574, 3893, 807.0], [100881, 5137, 653.0], [100905, 5212, 657.0], [100939, 5351, 353.0], [100952, 5412, 771.0], [100962, 5459, 739.0], [101749, 9604, 811.0], [101782, 9671, 809.0], [101843, 9915, 793.0], [101866, 9960, 808.0], [1, 1, 786.0], [2, 2, 786.0]]

Each entry in the list above has the information in the following format:

[event_id, ICD-9 code, rank]

Higher ranked events indicate a higher correlation score with the outcome variable. So the event that is highly correlated with his death is event_id 101866 which has the rank of 808 (highest amongst his recorded events) and ICD-9 code of 9960. The description for this code as recorded in the database is ‘CARDIOPULM RESUSCITA NOS’. This is a procedure performed on a patient with cardiac arrest to restore spontaneous blood

circulation and breathing. As can be understood this procedure indicates a life threatening disease as a cause and resulted in his mortality.

As pointed out earlier, the data used is not the one intended for this project. The results are not the diseases but are procedure that resulted in the death of a patient. But a close observation shows that the events that are positively and highly correlated with mortality are actually risky and life threatening and one's that are negatively correlated are not severely life threatening. Hence, it can be said that the model is working as expected. It is safe to assume that when this model is applied to the intended data, the results obtained will be relevant and accurate.

5. Evaluation:

To evaluate the model, cross validation method is used. 2-fold cross validation is used. The dataset is randomly sampled and is divided into two. Then the generated ranked list of events is compared with the actual ranked list to check if the events are ranked in the same order as in the list generated by the model. Following are the tables with the highest correlated 10 and least correlated 10 events, generated with one of the folds of the data:

event_id	ICD-9 Code (DESCRIPTION)	Correlation Score
101749	9604 (INSERT ENDOTRACHEAL TUBE)	0.2617
101783	9672 (CONTINUOUS INVASIVE MECH)	0.2228
101782	9671 (CONTINUOUS INVASIVE MECH)	0.19396

100574	3893 (VENOUS CATHETER NEC)	0.1867
101866	9960 (CARDIOPULM RESUSCITA NOS)	0.1846
100572	3891 (ARTERIAL CATHETERIZATION)	0.1634
100009	0017 (INFUSION OF VASOPRESSOR)	0.1398
101837	9907 (SERUM TRANSFUSION NEC)	00.1312
101835	9905 (PLATELET TRANSFUSION)	0.1251
101780	966 (EXT INFUS CONC NUTRITION)	0.1230

Table 3: Highest correlated 10 events with one of the folds of the data used for evaluation of the model

The highlighted features are the ones that have their rankings swapped when using the randomly sampled dataset. Otherwise, all the other features in the top 10 are ranked exactly the same as the full dataset. These variations could be because of the unbalanced distribution of the data with respect to the outcome variable. The following screenshot shows the distribution of these two events:

	event_id	value	count
	100574	N	5532
	100574	Y	1573
	101866	N	106
▶	101866	Y	205

Figure 7: Distribution of event-100574 and event-101866 with the outcome variable

The number examples available for the event 101866 are way too lesser than that of the event 100574. The random sampling could have resulted in further reduction in the number of examples available for the event 101866 and thus its correlation score.

Similarly in the table below with the least correlated 10 features, the one highlighted in red is completely missing from the top 10 list when using the randomly sampled dataset whereas, the one highlighted in yellow is a new addition.

event_id	ICD-9 Code (DESCRIPTION)	Correlation Score
101645	8853 (LT HEART ANGIOCARDIOGRAM)	-0.031415707099126874
100423	3521 (REPLACE AORT VALV-TISSUE)	-0.0312
101847	9920 (INJ/INF PLATELET INHIBIT)	-0.0318
101645	8853 (LT HEART ANGIOCARDIOGRAM)	-0.0347
100463	3613 ((AORTO)CORONARY BYPASS T)	-0.0489
100462	3612 ((AORTO)CORONARY BYPASS T)	-0.0519
100465	3615 (1 INT MAM-COR ART BYPASS)	-0.0806
100606	3961 (EXTRACORPOREAL CIRCULAT)	-0.0867
101881	9983 (OTHER PHOTOTHERAPY)	-0.0939
101078	640 (CIRCUMCISION)	-0.0958
101863	9955 (VACCINATION NEC)	-0.1689

Table 4: Least correlated 10 events with one of the folds of the data used for evaluation of the model

Apart from small variations in the ranks of the features, the algorithm is consistent with different sized datasets. The variations in the positively correlated procedures is

negligible as this variation is observed with the procedures with correlation almost equal to zero.

Apart from small variations in the correlation scores with respect to the events with very small number of observations, the model's performance is consistent with a varied sizes of the dataset.

6. Conclusion and Future Work:

A model has been learned that prioritizes events in a given patient's history. While there have been systems that summarized a patient's medical history on a single screen (e.g. LifeLines), none have used the data-driven approach. From a population of patients, the model learns the important events that are correlated with the mortality of a patient in the hospital. This knowledge is then applied on a given patient's history to prioritize his/her events. The model learned could successfully rank all the events in a dataset and use this "knowledge" to prioritize a given patient's health conditions. Apart from identifying events that are strongly correlated to mortality, the model can also identify procedures that are negatively correlated to mortality of a patient in the hospital. A negative correlation means that a procedure has resulted in a positive outcome in the population of data that it has been trained on. Having this information is useful as these suggest that the patients who have taken these treatments have benefitted. A clinician can use this information to make good treatment decisions.

This prioritization of events correlated with mortality (or in a general view, bad outcome) makes it possible for a doctor to utilize his appointment window effectively as the important events are highlighted. He can ask right questions and never miss significant details about a patient's health record.

This model is learned using data that is not historical in nature. But the same algorithm can be applied to re-learn the model for data which has historical medical data of patients. The plan is to develop three different models for the identified cohorts - chronic obstructive pulmonary disease (COPD), Heart Failure (HF), and Diabetes (DIAB). With IRB approval, historical data for patients admitted to UNC hospitals since 2008 was obtained. There are approximately 10000 patients in each of the cohorts. This data is collected from the UNC Clinical Data Warehouse (CDW). All the patients are adults (>17 years of age) who have both inpatient and outpatient medical data in the UNC CDW since 2008. This data is loaded into a MySQL database. However, due to access issues this data was not usable at the time of development of this model. For a given cohort, events that are closely related to the outcome variable will be ranked high and can be highlighted. Events and the outcome variables will be specific for each of the cohorts. When this dataset is used, the algorithm presented in this manuscript can be applied to learn new models.

Though the algorithm presented is generalizable, it is a very basic and a simple model. This can further be extended to identify clusters of commonly co-occurring features. This information can be used in the visualizations to group these co-occurring features. However, this study is limited to learning a model that can highlight events in a patient's history and any enhancements are for future implementation.

Visualization is also out of scope for this project. The ultimate goal is to summarize the individual patient's own medical data via interactive temporal visualization methods (e.g., "advanced timelines"). To be able to make a visualization which is interactive, the model has to be advanced. The temporal nature of the data has to

be dealt with properly to replicate the order of the events in a patient's medical history. To prevent any visual clutter, the visualization has to be properly abstracted into layers. These layers must be unfolded as required by the doctor. This kind of hierarchical visualization techniques for temporal data been implemented earlier by various studies like LifeLines and a similar approach can be taken. The highlighted features learnt from the existing populations of patients in a given cohort will be an advantageous addition in this study.

Once the project is ready, a feedback will be taken from the doctors at UNC Hospitals after they use the tool in their practice. Based on their feedback, new functionality can be incorporated (or the existing functionality could be fixed) to optimize the tool's usability.

Bibliography:

[1] Wilson, A. N. D. R. E. W. (1991). Consultation length in general practice: a review. *British Journal of General Practice*, *41*(344), 119-122.

[2] Mechanic, D., McAlpine, D. D., & Rosenthal, M. (2001). Are patients' office visits with physicians getting shorter?. *New England Journal of Medicine*,*344*(3), 198-204.

[3] Britten, N., Stevenson, F. A., Barry, C. A., Barber, N., & Bradley, C. P. (2000). Misunderstandings in prescribing decisions in general practice: qualitative study. *Bmj*, *320*(7233), 484-488.

[4] Barry, C. A., Bradley, C. P., Britten, N., Stevenson, F. A., & Barber, N. (2000). Patients' unvoiced agendas in general practice consultations: qualitative study.*Bmj*, *320*(7244), 1246-1250.

[5] Elder, N. C., & Dovey, S. M. (2002). Classification of medical errors and preventable adverse events in primary care: a synthesis of the literature. *The Journal of family practice*, (51), 927-32.

[6] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395-405.

[7] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, *309*(13), 1351-1352.

[8] Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). *Visual analytics: Scope and challenges* (pp. 76-90). Springer Berlin Heidelberg.

[9] Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. *Journal of medical systems*, *38*(4), 1-7.

[10] Aigner, W., Miksch, S., Müller, W., Schumann, H., & Tominski, C. (2007). Visualizing time-oriented data—a systematic view. *Computers & Graphics*,*31*(3), 401-409.

[11] Plaisant, C., Milash, B., Rose, A., Widoff, S., & Shneiderman, B. (1996, April). LifeLines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 221-227). ACM.

[12] Gallego, B., Walter, S. R., Day, R. O., Dunn, A. G., Sivaraman, V., Shah, N., ... & Coiera, E. (2015). Bringing cohort studies to the bedside: framework for a 'green button' to support clinical decision-making. *Journal of comparative effectiveness research*, *4*(3), 191-197.

[13] Longhurst, C. A., Harrington, R. A., & Shah, N. H. (2014). A 'green button' for using aggregate patient data at the point of care. *Health Affairs*, *33*(7), 1229-1235.

[14] Schork, N. J. (2015). Personalized medicine: time for one-person trials. *Nature*, *520*(7549), 609-611.

[15] Kurosaki, M., Hiramatsu, N., Sakamoto, M., Suzuki, Y., Iwasaki, M., Tamori, A., ... & Nakagawa, M. (2012). Data mining model using simple and readily available factors

could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. *Journal of hepatology*, *56*(3), 602-608.

[16] Plaisant, C., Lam, S., Shneiderman, B., Smith, M. S., Roseman, D., Marchand, G., ... & Rappaport, H. (2008, April). Searching electronic health records for temporal patterns in patient histories: a case study with microsoft amalga. In *AMIA*.

[17] McAullay, D., Williams, G., Chen, J., Jin, H., He, H., Sparks, R., & Kelman, C. (2005, January). A delivery framework for health data mining and analytics. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38* (pp. 381-387). Australian Computer Society, Inc..

[18] Toma, T., Bosman, R. J., Siebes, A., Peek, N., & Abu-Hanna, A. (2010). Learning predictive models that use pattern discovery—A bootstrap evaluative approach applied in organ functioning sequences. *Journal of biomedical informatics*, *43*(4), 578-586.

[19] Moskovitch, R., & Shahar, Y. (2009, November). Medical temporal-knowledge discovery via temporal abstraction. In *AMIA*.

[20] Wang, F., Lee, N., Hu, J., Sun, J., & Ebadollahi, S. (2012, August). Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 453-461). ACM.

[21] Schork, N. J. (2015). Personalized medicine: time for one-person trials.*Nature*, *520*(7549), 609-611.

[23] https://physionet.org/mimic2/mimic2_clinical_overview.shtml