USING MACHINE LEARNING METHODS TO IMPROVE HEALTHCARE DELIVERY IN DIABETES MANAGEMENT

By

Surya Bhaskar Ayyalasomayajula,

Bachelor of Science (Computer Science)

Delhi University, Delhi, 1987

Master of Business Analytics

Oklahoma State University, Stillwater, 2017

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2022

USING MACHINE LEARNING METHODS TO IMPROVE HEALTHCARE DELIVERY IN DIABETES MANAGEMENT

Dissertation Approved:

Dr. Dursun Delen

Dissertation Adviser

Dr. Rick Wilson

Dr. Chenzhang Bao

Dr. Rittika Shamsuddin (Outside member)

Name: SURYA BAHSKAR AYYALASOMAYAJULA

Date of Degree: JULY, 2022

Title of Study: AN ANALYTICS APPROACH TO IMPROVE HEALTHCARE DELIVERY IN DIABETES MANAGEMENT

Major Field: BUSINESS ADMINISTRATION

Abstract:

This dissertation includes three studies, all focusing on Analytics and Patients information for improving diabetes management, namely educating patients and early detection of comorbidities. In these studies, we develop topic modeling and artificial neural network to acquire, preprocess, model, and predict to minimize the burden on diabetic patients and healthcare providers.

The first essay explores the usage of Text Analytics, an unsupervised machine learning model, utilizing the vast data available on social media to improve diabetes education of the patients in managing the condition. Mainly we show the applicability of topic modeling to identify the gaps in diabetes education content and the information and knowledge needs of the patients. While traditional methods of the content decision were based on a group of experts' contributions, our proposed methodology considers the questions raised on social forums for support to extend the education content.

The second essay implements Deep Neural Networks on EHR data to assist the clinicians in rank ordering the potential comorbidities that the specific patient may develop in the future. This essay helps prioritize regular screening for comorbidities and rationalize the screening process to improve adherence and effectiveness. Our model prediction helps identify diabetic retinopathy and nephropathy patients with very high precision compared to other traditional methods. Essays 1 and 2 focus on Data Analytics as a research tool for managing a chronic disease in the healthcare environment.

The third essay goes through the challenges and best practices of data preprocessing for Analytics studies in healthcare. This study explores the standard preprocessing methodologies and their impact in the case of healthcare data analytics. Highlights the relevant modifications and adaptations to the standards CRISP_DM process. The suggestions are based on past research and the experience obtained in the projects discussed earlier in the thesis.

Overall, the dissertation highlights the importance of data analytics in healthcare for better managing and diagnosing chronic diseases. It unfolds the economic value of implementing state-of-the-art IT methods in healthcare, where EHR & IT are predominantly costly and difficult to implement. The dissertation covers ANN and text mining implementation for diabetes management.

TABLE OF CONTENTS

Chapter Page No.

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Healthcare is the organized provision of medical care to individuals or a community. Traditionally an analytically driven approach is applied to in healthcare organizations to optimize the resources. Since, time immemorial healthcare is resource scarce, especially trained physicians and other medical personnel. The physicians, based on their experience and experimental results, treat new patients. Gregor Mendel's theory on genetics is an excellent case of analytics based on empirical data. Recent World Health Organization's (WHO) report estimate there are 14 physicians per 10,000 population and the physicians are mostly concentrated in urban areas. Instead of increasing the number of physicians, their services should be delivered to the rural and remote population too. Telehealth and analytics do this job, by providing services remotely and by early diagnosis or by providing preventive treatment by using analytics.

Hassan (2019) examined the implications of the complex origins of data analytics and data science for the IS field, specifically on how different discourses impact future research and practice. With the development of algorithms and availability of data, Healthcare and businesses were the early implementers of Data Analytics. Analytics has been implemented by commercial organizations to maximize profit and Not for Profit institutions had incorporated analytics in their operations to maximize the service provided. Healthcare, Agriculture, Sustainability, and Climate change are some of the areas where analytics are implemented, and they need further implementation.

Interoperability and standardization of organizational resources in healthcare is essential for faster and better care of the patient especially during emergencies, this is the primary reason the US government has subsidized the EHR adoption. Lehne, Luijten, Vom Felde Genannt Imbusch, and Thun (2019) has studied and elaborated positively on this phenomenon by presenting a systematic review of articles published on interoperability.

1.1 MACHINE LEARNING AND DEEP LEARNING

Machine learning is a method of data analysis that automates analytical model building eliminating the need of human intervention and inputs. It is a branch of artificial intelligence based on the idea that systems can learn from available datasets, identify patterns and make decisions with minimal expert human intervention. Because of new computing technologies such as graphical processing units (GPU) and advanced algorithms, machine learning today is not like what it used to be in the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence devised methods and algorithms for computers to learn from data. The iterative aspect of machine learning is critical because as models are exposed to additional datasets, they are able to adapt without human intervention. They learn from computations to produce reliable, repeatable decisions and results. It’s a science that has been designed by mathematicians and is not new – but one that has gained fresh importance and recognition with advancement and availability of fast and ample computing power. While many machine learning algorithms have been developed an in use for a long time, the ability to automate the application of complex mathematical calculations to big data – over and over, faster and faster – is a current development.

Deep learning is a type of machine learning and artificial intelligence (AI) tries to imitate one of the ways humans acquire certain types of knowledge. Deep learning is a critical component of data science, along with statistics and predictive modeling. This is extremely beneficial to data scientists who are involved in collecting, analyzing and interpreting big data; deep learning makes this process faster, easier and feasible.

At its simplest form, deep learning can be thought of as a way to automation of analytics. While traditional machine learning algorithms are best suited to identify linear relations, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction thus achieving the improved performance.

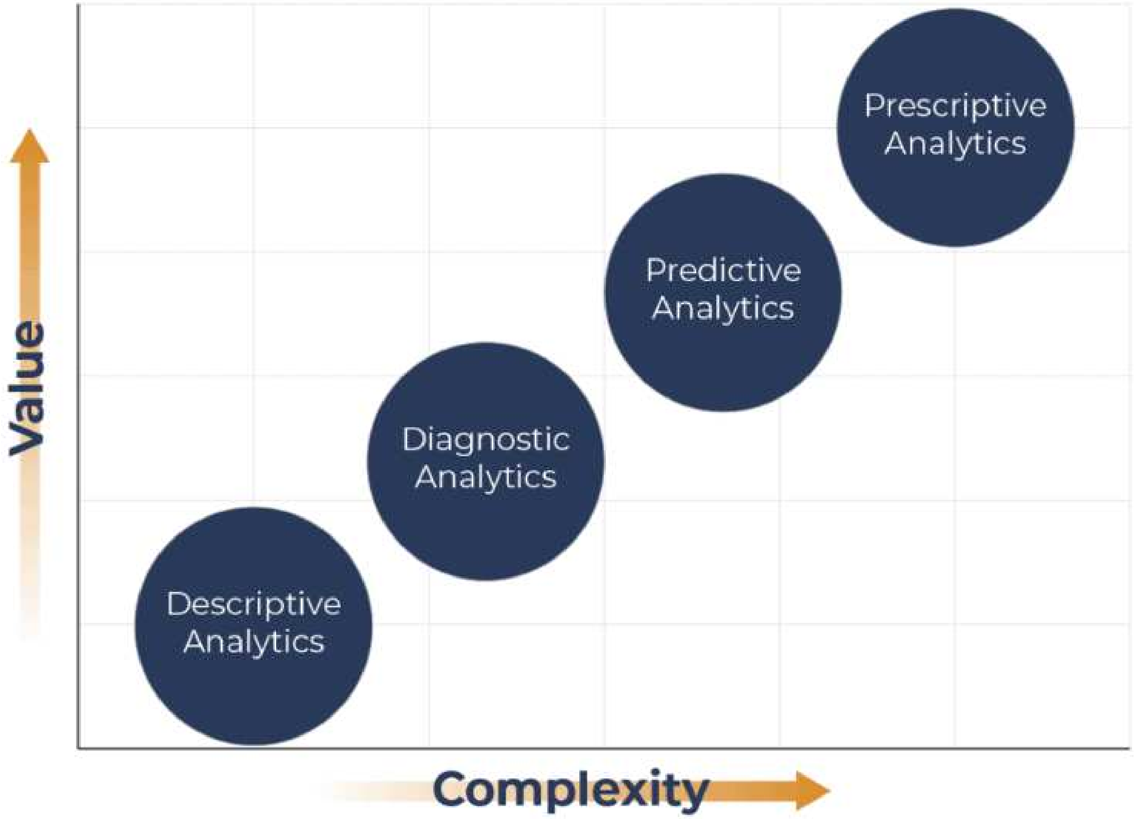


Figure 1.1 Data analytics progression and classification

Data analytics can be classified into four sub-divisions Descriptive, Diagnostic, Predictive, and Prescriptive. As shown in figure 1.1, they progressively increase in value and complexity. Organizations also follow this path while adopting analytics in their operations.

Descriptive analytics, as the name implies, describes what has already happened so that we can understand and probably learn some lessons for the future. Diagnostic analytics is diagnosing or analyzing why did the event happen. We want to know the causation of the events so that we may control the events or predict in the future. Predictive analytics is predicting the future similar to an astrologer but based on the facts and scientific and sound models and algorithms. The prescriptive realm is to intervene and modify the outcome as predicted by predictive analytics. The objective of the analytics can always be categorized in the above four echelons.

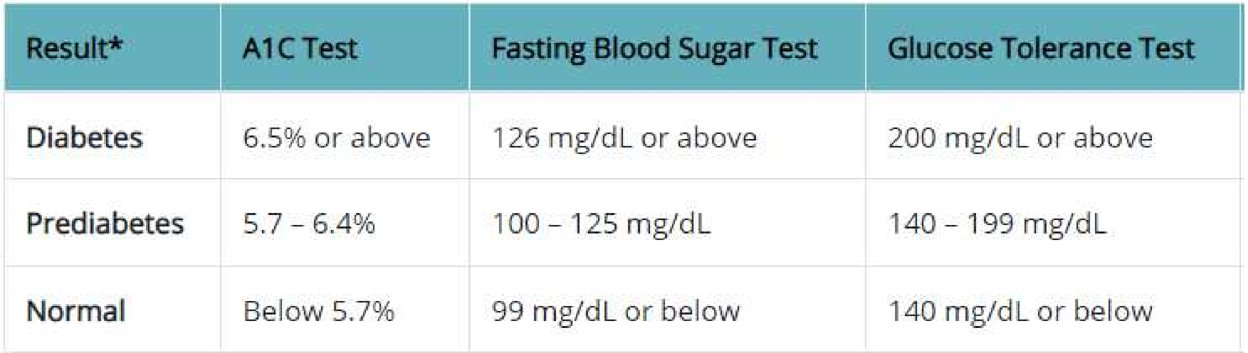
1.2 HEALTHCARE MANAGEMENT

Healthcare, is the maintenance, and improvement of physical and mental health via the prevention, diagnosis, treatment, amelioration, or cure of disease, illness, injury in human beings. One of the major diseases which is affecting the humanity is diabetes. Approximately 34.2 million people or 10.2% population in the United States have diabetes, up from the previous estimate of 26 million in 2010, additionally 33% patients with diabetes doesn't know he or she has it (Centers for Disease Control Prevention & Services, 2020). As per the same report another 88 million adults are prediabetic, which may lead to diabetes in future if not taken care.

From an economic point of view, US spent $3.8 trillion amounting to 17.7% of its GDP on health care, which was 4.6% more than the previous year (Centers for Medicare & Medicaid Services, 2021). As per The Commonwealth Fund, the United States spends the highest GDP on healthcare among High-Income countries, but as per the outcomes, it stands 11th in the order, much below many other nations which spend less. The implementation of Data Analytics has brought higher efficiency in resource usage and delivery time in almost all the industry sectors and we estimate that the same would be case in Healthcare and diabetes management too.

The clinical definition of diabetes mellitus and its effects is "Diabetes is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both. The chronic hyperglycemia of diabetes is associated with long-term damage, dysfunction, and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels" (AMERICAN DIABETES ASSOCIATION, 2004).

Table 1.1 Classification and diagnosis of diabetes



Result*	A1C Test	Fasting Blood Sugar Test	Glucose Tolerance Test
Diabetes	6.5% or above	126 mg/dL or above	200 mg/dL or above
Prediabetes	5.7 – 6.4%	100 – 125 mg/dL	140 – 199 mg/dL
Normal	Below 5.7%	99 mg/dL or below	140 mg/dL or below

Source: Adapted from American Diabetes Association. *Diabetes Care.* 2016;39(1): S14–S20, tables 2.1, 2.3.

Diabetes is a condition where the amount of sugar present in the blood is not normal. The sugar level in blood is checked in two units based on the test performed, and it can be either Milligram per deciliter or Millimoles per liter. Less than 140mg/dL is considered normal, and > 200mg/dL is deemed to be diabetic. The A1C test gives the average sugar level of the last 3 months in the blood. Fasting Blood Sugar test, as the name implies your blood is tested after you fast for 8 hours or more, normally the blood is drawn in the morning before breakfast. For the Glucose Tolerance test, after fasting you are made to drink glucose and tested after 1, 2, 3 hours. If any one of the tests comes positive than you are considered diabetic.

There is no cure for diabetes, and it is the 7th most leading cause for mortality rate in US. It needs to be actively managed so that other comorbidities may not develop. The management strategy is to keep the sugar level within range by using medicines in the form of pills and, in some cases injecting insulin, changes in diet, regular exercise, and physical activity. Patients need to monitor their sugar levels by pricking their fingers using a lancet and then using a testing strip and monitor. Continuous Glucose Monitors (CGM) are another option where frequent and regular monitoring is needed. It is a chronic disease lasting lifelong, and there are infrequent exceptions. As there is no cure. The target of management is to keep the blood sugar levels within range and avoid the risk of getting other comorbidities associated with diabetes. Abrahams, Jiao, Wang, and Fan (2012) employed text mining to identify and prioritize defects in automobiles based on the discussions on a popular social medium used by vehicle enthusiasts.

Brahma, Goldberg, Zaman, and Aloiso (2021) used machine learning and Text Analytic techniques to predict likely delays in mortgage origination so that extra emphasis can be given to those cases to minimize rejection or delay. Chen, Zheng, Xu, Liu, and Wang (2018) designed a text analytics framework to assess secondhand sellers' reputation online. For online retail brands Ibrahim and Wang (2019) identified the customers' primary topics of concern that are shared among Twitter users.Wang and Xu (2018) proposed a novel deep learning model for vehicle insurance fraud detection that uses the novel Latent Dirichlet Allocation (LDA)-based text analytics on descriptive textual portion of the insurance claim reports. Hassan Zadeh and Jeyaraj (2018) using text mining analysis have examined the extent to which the business strategies and social media strategies of organizations are aligned based on twitter feeds and annual 10-K reports.

1.3 DIABETES MANAGEMENT

Diabetes can be grouped into three subgroups. Type 1 diabetes, also known as juvenile diabetes or insulin-dependent diabetes, the pancreas produces negligible or no insulin. Type 2 diabetes (DT2M), pancreas does not produce enough insulin and cells are not able to utilize available insulin properly and consume sugar inefficiently. Gestational diabetes is a type of diabetes that can develop during pregnancy in women who don't already have diabetes.

Diabetes can be managed by education, medication, and lifestyle changes of proper diet, physical activity, and weight management. This is achieved by medication (Pills and, if necessary, insulin injections or insulin pumps). Taking a healthy and balanced diet, being physically active, and exercising in moderation are very important.

As per American Diabetes Association (ADA), the diabetes management plan is not simple, an ongoing process, and needs adjustment on a regular basis (American Diabetes, 2019). The patients can enroll in classes offered by the hospital, local ADA chapters, or other community wellness centers. They also avail one on one sessions with the Certified Diabetes Educator. The diabetes education would cover nutrition, insulin dosage, adequate exercise, monitoring blood glucose and urine ketone.

Eriksson and Lindgärde in their study, published over three decades ago, has shown that DT2M is preventable or delayable by lifestyle interventions (exercise and proper diet) in persons at risk of the disease (Hawley & Gibala, 2012).

1.4 DIABETES COMORBIDITY

Generally, diabetes leads to many other complications, and it is vital to monitor and screen for those complications regularly. According to the CDC, the hospitalized population had the following comorbidities and patient populations as listed in Table 1.2.

Table 1.2 Comorbidities and patient populations among the hospitalized

Risk Factor	Number in thousands	Crude rate per 1,000 (95% CI)
Diabetes as any listed diagnosis	**7,833**	**339.0 (317.6–360.4)**
Major cardiovascular disease	1,740	75.3 (70.4–80.2)

Ischemic heart disease	438	18.9 (17.6–20.2)
Stroke	313	13.6 (12.7–14.5)
Lower-extremity amputation	130	5.6 (5.3–6.0)
Hyperglycemic crisis	209	9.1 (8.5–9.6)
Diabetic ketoacidosis	188	8.1 (7.6–8.7)
Hyperosmolar hyperglycemic syndrome	21	0.9 (0.85–1.0)
Hypoglycemia	57	2.5 (2.3–2.6)

To prevent and monitor these and other comorbidities, the ADA recommends laboratory evaluations, fundoscopic examination by an eye specialist, Thyroid palpitations, and comprehensive foot examination to be conducted every year.

Clinicians and researchers recognize the need to reorganize health systems to better serve diabetic patients, but health system have focused narrowly on diabetes management and few other conditions only. Diabetic patients often have comorbid conditions and other health problems as well, which increases their struggle to manage their health (Piette & Kerr, 2006).

1.5 SCOPE OF THE DISSERTATION

This multidisciplinary dissertation draws upon healthcare, data analytics, and information systems. The problem of chronic diseases (Diabetes) in healthcare is studied by applying data analytics tools and techniques utilizing data stored in healthcare information systems. Figure 1.1 presents a Venn diagram describing the scope of this dissertation at the intersection of three areas: healthcare, data analytics, and information systems.

In healthcare, we enhance diabetes management by providing the framework to improve diabetes education and predict the probability of comorbidities. Our study applies supervised and unsupervised machine learning algorithms to study diabetes. Chronic diseases must be addressed by prevention, management, and temporal delaying of further complications and comorbidities.

The models and their results are expected to improve the screening adherence and contents of diabetes education management.

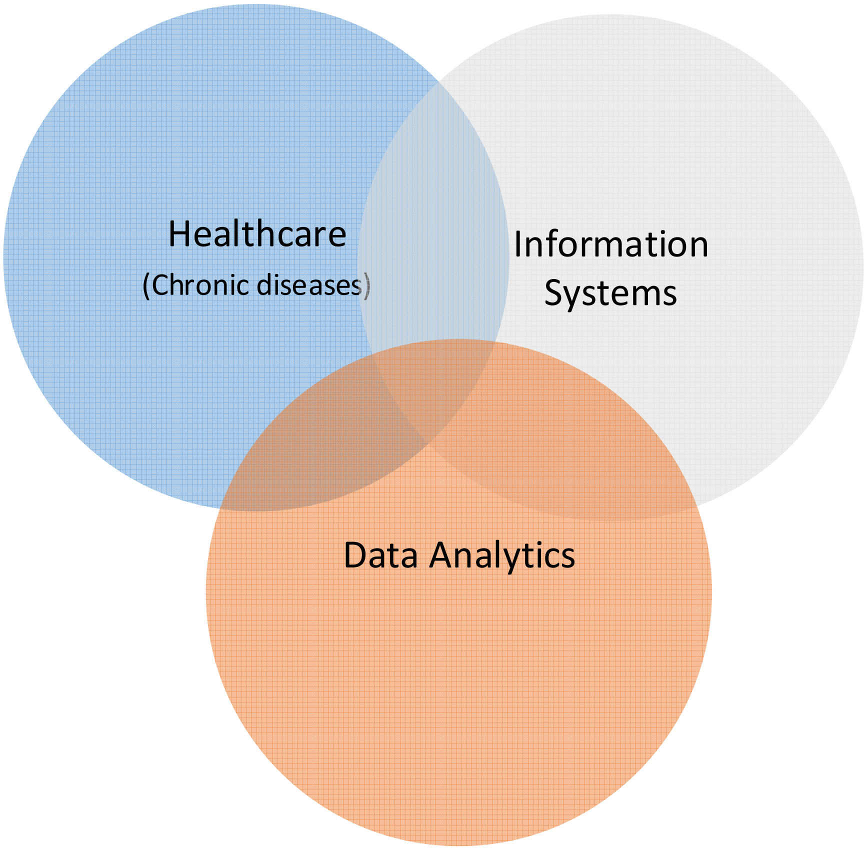


Figure 1.2 Venn diagram explaining scope of the dissertation

1.6 CURRENT STUDY

Prior studies in Data Analytics have examined how various methods can be utilized to detect the diseases based on analytical methods (Piri, Delen, Liu, & Zolbanin, 2017) (Edward Choi, Bahadori, Schuetz, Stewart, & Sun, 2016), diagnose cancers (Saba & Technique, 2021), reading X-rays reports. On the economic side, empirical studies and CDC reports evaluate the cost of various diseases on the government and public and the direct and indirect costs to the economy. Dutta, Batabyal, Basu, and Acton (2020) developed an efficient convolutional neural network for coronary heart disease prediction. Despite the growing literature, existing studies have not fully considered Data Analytics for the assistance of healthcare personals and temporal displacement of care in the case of chronic diseases, specifically diabetes.

Many studies on preprocessing the data for machine learning and deep learning in various domains. But the challenges of data preprocessing are unique in Healthcare and are not widely studied. We address these issues in the three essays. Generally, diabetes leads to many other complications, and it is vital to monitor and screen for those complications regularly.

Text Analytics and topic modeling are well-established methods in many fields where there is lots of text data available. This methodology has been used in the EHR notes and for HIPPA deidentification. One of the research study is De-identification of clinical notes via recurrent neural network and conditional random field (Z. Liu, Tang, Wang, & Chen, 2017). In many other domains, it has been used for the social media content also. As healthcare is a very sensitive and regulated industry, text analytics is not implemented widely. In the first essay, we want to identify the gaps in Diabetic Education, which is designed by the medical experts and unable to answer all the doubts and questions of the patients. The topic modeling portion of the study can be classified as descriptive analytics.

Developing comorbidities in the case of chronic diseases is a critical phenomenon. To address this issue, screening for related comorbidities is designed by the experts. The multitude of screening, accessibility and time constraints cause the patients to not adhere to the screening schedule. The list of comorbidities is considerable, and consequently, the screenings are also many and at different time intervals and specialty locations. Using deep learning on the EHR data, we can rank order the probability of comorbidities and highlight the most needed screenings to be done without fail. The second essay implements the deep learning framework in the case of diabetes for rank-ordering the likelihood of comorbidity. This fall under the predictive analytics echelon as we are predicting the probability of the comorbidities.

During these projects, we faced many issues related to healthcare data preprocessing for implementing machine learning models. Our study of previous literature sheds light on these issues, which are common and unique to healthcare. The third essay highlights the data preprocessing issues faced by other researchers with healthcare data and the best practices to resolve these issues.

The three essays will add to the Healthcare Analytics literature in several ways. Essay 1 highlights the important role of Text Analytics and Social Media in improving health care patient education. Essay 2 unfolds the deep learning and its implementation in EHR-based healthcare decision support systems. Essay 3 guides preprocessing the data for machine learning and deep learning in the healthcare domain. While the three essays cover a wide range of healthcare improvement methods, they all are inherently linked to Machine / Deep learning and predictive analytics.

CHAPTER II

ESSAY I: CONTINUALLY IMPROVING DIABETES EDUCATION BASED ON PATIENT'S SOCIAL MEDIA INTERACTIONS USING TEXT ANALYTICS

ABSTRACT

The pressing need for diabetes patients' education include addressing their concerns about different types of medicines, insulin injections methods and dosage, and usage of glucose monitoring devices. This paper employs text mining to analyze open forum discussions posted on a popular online diabetes support group frequently visited by diabetes patients organized and maintained by diabetes.org. The findings suggest that the expert advice and education based on research and clinical trials are not sufficient to address diabetes patient concerns and inquiries. The current study proposes, describes and evaluates a new process, and a decision support system to properly discover diabetes patient concerns and their prioritization. In broad context our findings also provide insights into how text analytics can be used to improve diabetes education for patients specifically, and perhaps for patients with any other chronic diseases.

2.1 INTRODUCTION

Diabetes is a chronic disease without a cure, and the only viable option is to manage it. In the management of diabetes, the patient and their family members play an important role, along with the physician and other support staff. Healthcare professionals also rely on the proactive involvement of the patient's domestic caregivers or social structure supporters in managing this incurable disease. Hence, disseminating knowledge and addressing the concerns of these groups of persons is essential for proper management. Shaw et al. (2020) have studied and identified that Self-monitoring diabetes with multiple mobile health devices and uploading the data seamlessly into EHR systems will be the future trend in managing diabetes.

Diabetes Self-Management Education and Support (DSMES) services help people with diabetes learn how to take the best care of themselves. DSMES services include a health care team who will teach how to stay healthy and make the knowledge gained a regular part of the diabetic patients' life. DSMES aims to learn the skills to Eat healthily, be physically active, check blood sugar (glucose), Take medicine properly, Cope with the emotional side of diabetes and reduce your risk of other health problems.

Medical professionals use a lot of medical jargon and technical terms while interacting with patients. They try to explain and provide simple explanations for the conditions, but the patients cannot decipher the contents thoroughly. Due to the limited interaction time, the patients try to find an alternate source of information and clarification. A. N. Miller et al. (2022) have found that just 50% of jargon terms were explained, and the male medical professionals use 50% more jargon leaving the patients perplexed and ill-informed.

2.1.1 AN OVERVIEW OF TEXT ANALYTICS

A broad stream of text analytic algorithms and methodologies are developed and implemented in the scientific and academic research community. The academic and social interaction text follows the language and grammar structure which is difficult for the computer to understand and conceptualize. Natural language processing (NLP) is a stream of algorithms and models, where the sentences are broken into the parts of speech (POS), i.e., noun, verb, subject, pronoun, tense, etc., and then studied. Grammarly is an application that utilizes these algorithms, informs us of the deficiencies in the text, and suggests improvement.

Processing textual data, specifically interviews and notes, was done via coding in qualitative research. It comprises categorizing and thematically sorting the text and providing an organized platform for constructing meaning and getting a summary. Qualitative research employs data coding methods that reveal themes embedded in the data. Coding is a critical structural operation in qualitative research, enabling data analysis and successive steps to serve the purpose of the study. Topic modeling in text analytics is a machine-learning algorithm to achieve the same: to recognize patterns and get insights without knowing the language structure and contents.

This methodology can be applied to any language and is widely known algorithm of text analytics. The first step in text analytics is converting the unstructured text into a structured form to store and use in the algorithms. This is achieved by converting the text into a multidimensional numerical matrix representation. The simplest form is the consideration of the document as a bag of words (BOW), and a dictionary with the frequency of occurrences is generated. The most straightforward visual representation of the dictionary is a word cloud.

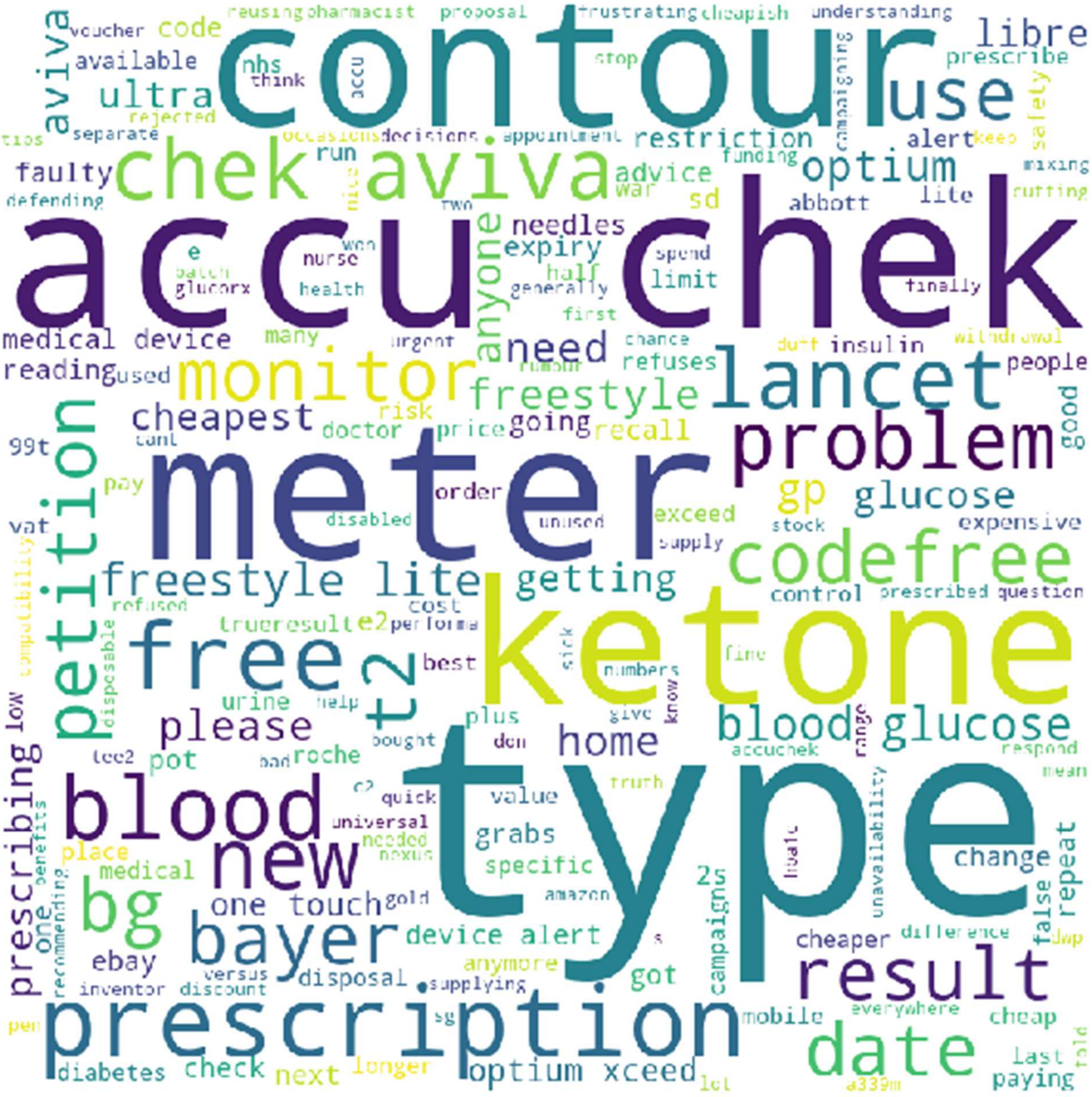


Figure 2.1 Word cloud of diabetic threads

Stanford University researchers have studied a corpus of documents and assigned sentiment scores, both positive and negative, for most words in the English language. This corpus is available as SentiNet library in python. We can assign a positive or negative score by aggregating the scores of all the words in a given document (review/tweet) based on the SentiNet corpus. This is a sentiment analysis algorithm, researchers usually modify the algorithm based on context and textual sources to get better insights.

Usually, documents have a title assigned by the author. In some cases, the title may not be a correct representation of the document contents. In many other cases, the title or summary might be missing or not assigned, as in the case of tweets and some social media posts. There are two algorithmic families to cluster the documents based on the topic of the document: BOW, and the other is Term Frequency / Inverse Document Frequency (TF/IDF). In both cases, we have to specify the number of topics the corpus of documents is about. A better and more complex algorithm is Latent Dirichlet Allocation (LDA). In this algorithm, each document is considered to be about multiple topics with different weights.

2.1.2 TEXT ANALYTICS GENERAL CONTEXT

Feuerriegel and Gordon (2018) used a Long-term stock index forecasting based on text mining of regulatory disclosures. (Gruss, Abrahams, Fan, & Wang, 2018) have developed a system by which the numerical information and scale are extracted from text documents. We usually remove the numbers from the document while doing the text analytics. Lee and Zhao (2020) developed a model to predict the helpfulness of online reviews specific to a business entity. An investigation of Linguistic Inquiry and Word Count (LIWC), which calculates the presence of >80 language dimensions in-text samples, and permits the construction of custom dictionaries, demonstrates the use of LIWC to ensure better problem/model fit within the context of selecting a decision support tool (McHaney, Tako, & Robinson, 2018). Tao, Deokar, and Deshmukh (2018) have analyzed forward-looking statements in initial public offering prospectuses using a text analytics approach to identify the impact on IPO valuation. Yuan, Lau, and Xu (2016) have identified the determinants of crowdfunding success using a semantic text analytics approach.

2.1.3 TEXT ANALYTICS MEDICAL NOTES

Ayyar and IV (2016) and Baumel, Nassour-Kassis, Cohen, Elhadad, and Elhadad (2017) has developed models utilizing the medical notes to assign ICD codes at the discharge time, enhancing the accuracy of billing for the hospital. Swiss Variant interpretation Platform for Oncology has utilized Text -Mining to develop services for annotation of MEDLINE and PMC articles (Caucheteur et al., 2020). Abidi, Singh, and Christie (2017) has used text analytics to automatically fill a medical case report from unstructured referral letters by reducing the time to fill the case report. Cassim, Mapundu, Olago, George, and Glencross (2019) have shown the ability of predictive analytics and text mining techniques to extract prognostic information from reports with semi-structured narrative text that is not easily analyzable to guide patient management.

2.1.4 TEXT ANALYTICS MEDICAL OTHERS

Chatterjee, Deng, Liu, Shan, and Jiao (2018) have developed a more sophisticated Text Analytics model to segregate facts versus opinions in case of tweets which would assist customer service requests. Instead of text analytics, Christopoulou, Tran, Sahu, Miwa, and Ananiadou (2019) have utilized state-of-the-art named-entity recognition (NER) models based on bidirectional long short-term memory (BiLSTM) networks to identify Adverse Drug Events in standard EHR notes.

2.1.2 DIABETES

The increase in the number of diabetic patients and the negative impact on the economy and health have alerted governments and health organizations to provide education outside the clinical setting to prevent and manage diabetes. Still, many patients feel the inadequacy of these measures and turn to social media and experienced patients for advice and assistance. Web sites and books have documented the strides in diabetic management and informed the best practices to manage the disease. Still, the patients' needs are not entirely met. The educational materials (books, videos, and websites) would be much more effective if they could address most of the questions and doubts the patients have. However, the list of online support groups and numerous questions posed in these forums highlight the inadequacy of the education material. We suggest that comprehending the topics discussed in these forums would assist the Healthcare professional and government agencies in updating their offerings of educational content and improving meeting patient needs.

2.2 DIABETES EDUCATION

Diabetes is a chronic and non-curable disease. Hence patients are trained to medicate and monitor themselves at home with regular (monthly or quarterly) hospital visits, thus avoiding costly hospital visits and admissions. The patient has to perform blood tests and inject insulin if necessary regularly. In the case of other diseases or conditions, these actions (Blood testing and injecting) are performed by trained and licensed medical professionals in a clinical setting only. However, in the case of diabetes, to reduce cost and provide a better experience, the patient or the domestic caregiver is trained to monitor blood sugar levels and inject insulin at home without medical support or supervision.

Educational training for the management of diabetes is essential as it is a core component of disease management. The physicians and caregivers accomplish training by providing hands-on training, creating websites with video tutorials and articles, and publishing books and broachers written by expert medical professionals and associations. Mayo clinic, diabetes.org.uk, American Diabetes Association (ADA), and the Center for disease control and prevention (CDC) have detailed and well-informed websites for diabetes patients. Guides to diabetes are written by Mayo clinic, ADA, The medical Library association, and other expert medical professionals. There are diabetic support groups online with thousands of members across nations. These members ask lots of questions and seek clarification, answered or discussed by the experienced users and medical professionals. Such support groups are typical for diabetes, cancers, and other chronic diseases.

Diabetes UK was founded in 1934 as “The Diabetic Association” by the famous authors H. G. Wells and Dr. R. D. Lawrence. It supports diabetes patients and disseminates knowledge and care information, and it is a research charity. It has funded more than $66 million in the last decade. This organization changed with the times and adopted the latest technology, and its web address is Diabetes.org.uk. The large number of threads on this forum shows that the training and diabetes literature is not addressing all the questions faced by diabetes patients. Hawley and Gibala (2012) have developed a novel and time-efficient interval training for diseases like diabetes which afflict populations at risk of inactivity-related diseases.

2.3 METHODOLOGY

We followed the best practices as advocated by the CRISP-DM process management. We have gone through the initial five steps: Business Understanding, Data Understanding, Data Preparation, and Modeling Deployment. We intend to execute the last step of Evaluation in partnership with a publisher or medical association. In the previous sections, we have described the Business Understanding that is the context of the problem.

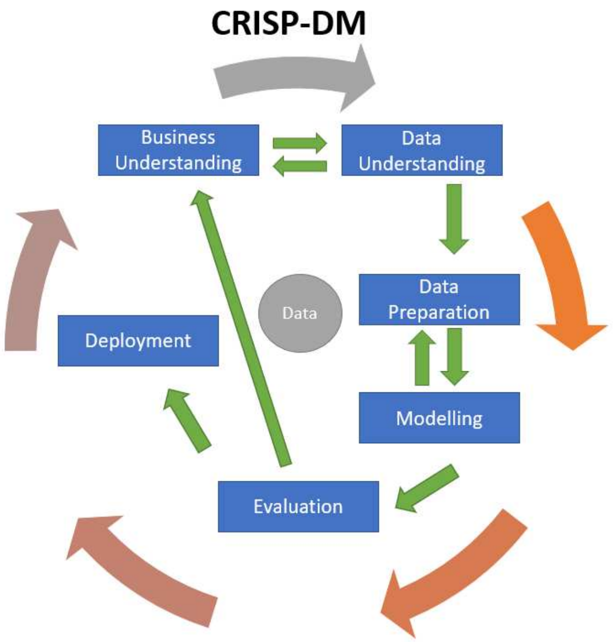
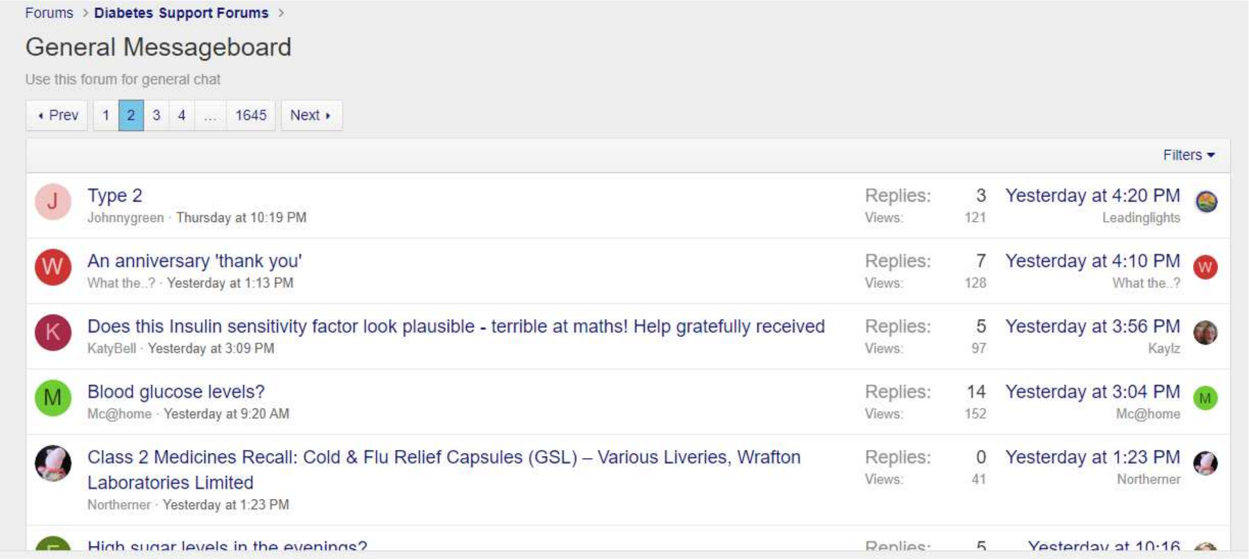


Figure 2.2 CRISP-DM methodology for data analytics

2.3.1 DATA ACQUISITION

The discussion forum at Diabetes.org.uk has 30 thousand threads related to diabetes only. This forum has more than 23,000 active users who have posted more than one Million queries and responses.



Forums > **Diabetes Support Forums** >

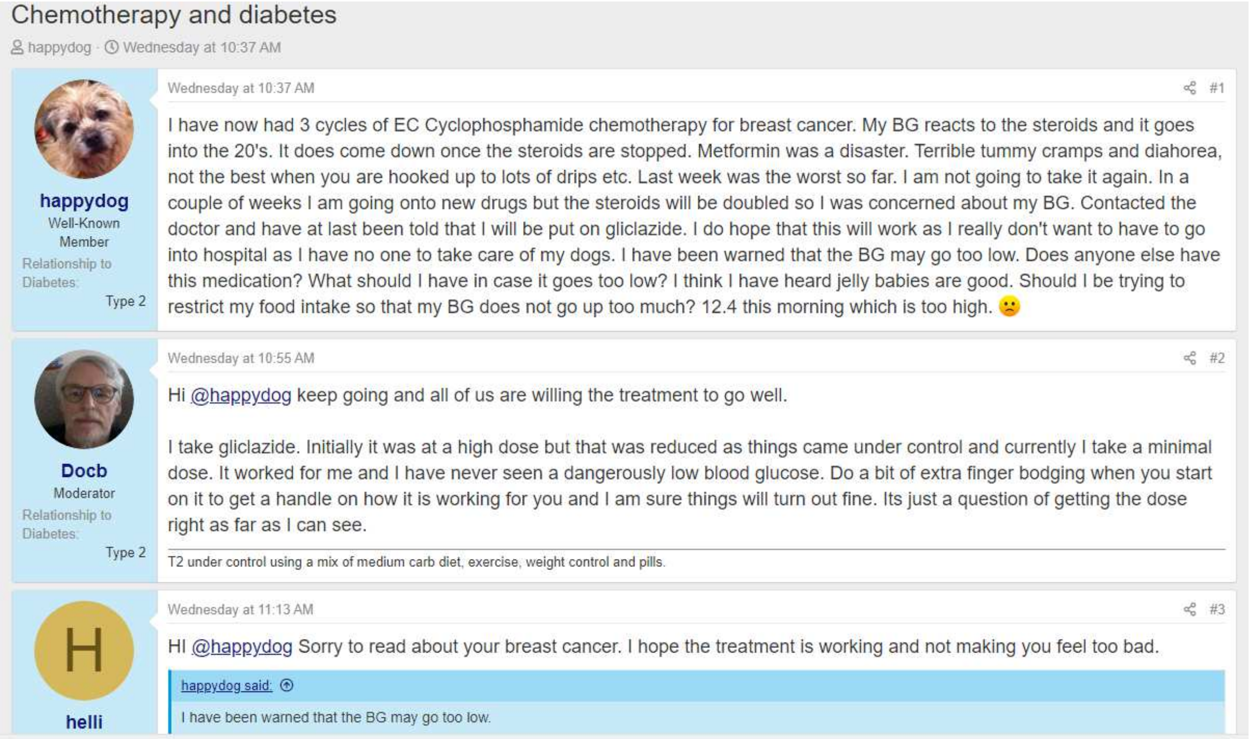
General Messageboard

Use this forum for general chat

J	Type 2 Johnnygreen · Thursday at 10:19 PM	Replies: 3 Views: 121	Yesterday at 4:20 PM Leadinglights	
W	An anniversary 'thank you' What the..? · Yesterday at 1:13 PM	Replies: 7 Views: 128	Yesterday at 4:10 PM What the..?	W
K	Does this Insulin sensitivity factor look plausible - terrible at maths! Help gratefully received KatyBell · Yesterday at 3:09 PM	Replies: 5 Views: 97	Yesterday at 3:56 PM Kaylz	
M	Blood glucose levels? Mc@home · Yesterday at 9:20 AM	Replies: 14 Views: 152	Yesterday at 3:04 PM Mc@home	M
	Class 2 Medicines Recall: Cold & Flu Relief Capsules (GSL) – Various Liveries, Wrafton Laboratories Limited Northerner · Yesterday at 1:23 PM	Replies: 0 Views: 41	Yesterday at 1:23 PM Northerner	
	High sugar levels in the evenings?	Replies: 5	Yesterday at 10:16	

Figure 2.3 Diabetes.Org.Ok Message Board

The queries are called threads, and each thread has a title and a detailed reply by the participating users in response to the communication held to date. Some responses are specifically in reply to a particular portion of the conversation. In such cases, the original textual portion is highlighted and quoted. Each reply has the user name, date and time, and the reply sequence number. The members in this forum discuss diabetes and have stated their diabetes relationship in the user profile. We cannot infer or verify the quantity, quality, and source of diabetes education that the users have obtained. Hence, we have to use caution and subjective judgment when giving weightage to the results based on these posts.



Chemotherapy and diabetes

happydog · Wednesday at 10:37 AM

happydog
Well-Known Member
Relationship to Diabetes: Type 2

Wednesday at 10:37 AM #1

I have now had 3 cycles of EC Cyclophosphamide chemotherapy for breast cancer. My BG reacts to the steroids and it goes into the 20's. It does come down once the steroids are stopped. Metformin was a disaster. Terrible tummy cramps and diahorea, not the best when you are hooked up to lots of drips etc. Last week was the worst so far. I am not going to take it again. In a couple of weeks I am going onto new drugs but the steroids will be doubled so I was concerned about my BG. Contacted the doctor and have at last been told that I will be put on gliclazide. I do hope that this will work as I really don't want to have to go into hospital as I have no one to take care of my dogs. I have been warned that the BG may go too low. Does anyone else have this medication? What should I have in case it goes too low? I think I have heard jelly babies are good. Should I be trying to restrict my food intake so that my BG does not go up too much? 12.4 this morning which is too high. ☹

Docb
Moderator
Relationship to Diabetes: Type 2

Wednesday at 10:55 AM #2

Hi @happydog keep going and all of us are willing the treatment to go well.

I take gliclazide. Initially it was at a high dose but that was reduced as things came under control and currently I take a minimal dose. It worked for me and I have never seen a dangerously low blood glucose. Do a bit of extra finger bodging when you start on it to get a handle on how it is working for you and I am sure things will turn out fine. Its just a question of getting the dose right as far as I can see.

T2 under control using a mix of medium carb diet, exercise, weight control and pills.

H
helli

Wednesday at 11:13 AM #3

HI @happydog Sorry to read about your breast cancer. I hope the treatment is working and not making you feel too bad.

> happydog said: ⊕
>
> I have been warned that the BG may go too low.

Figure 2.4 Detailed Posts At Diabetes.Org.Ok

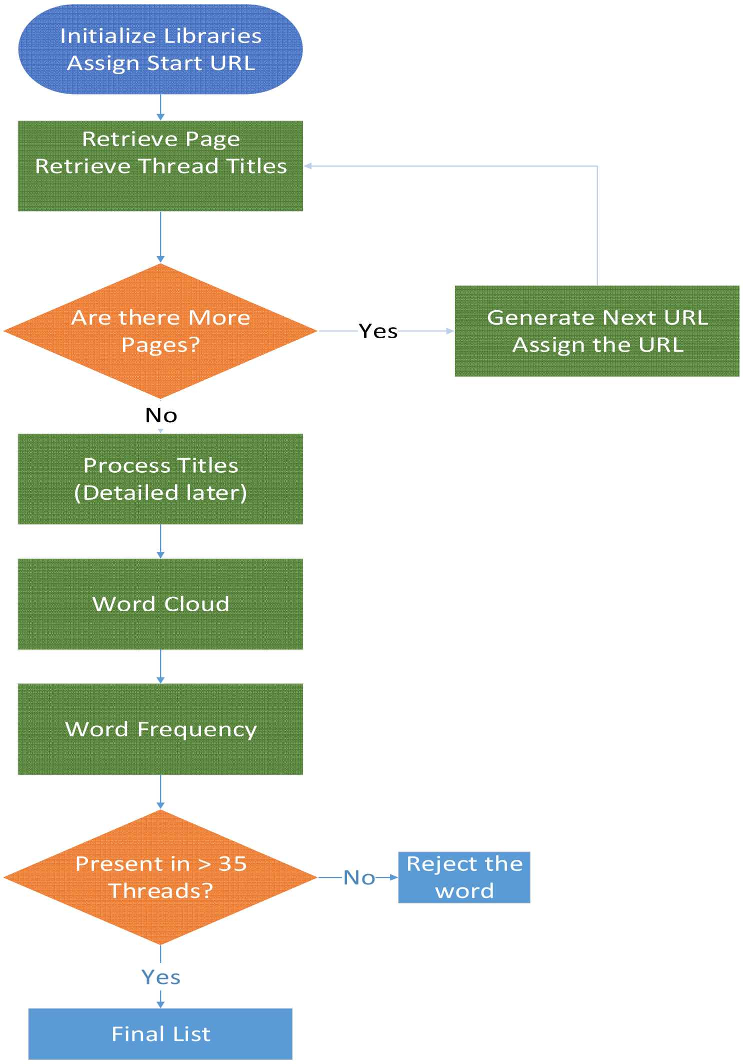
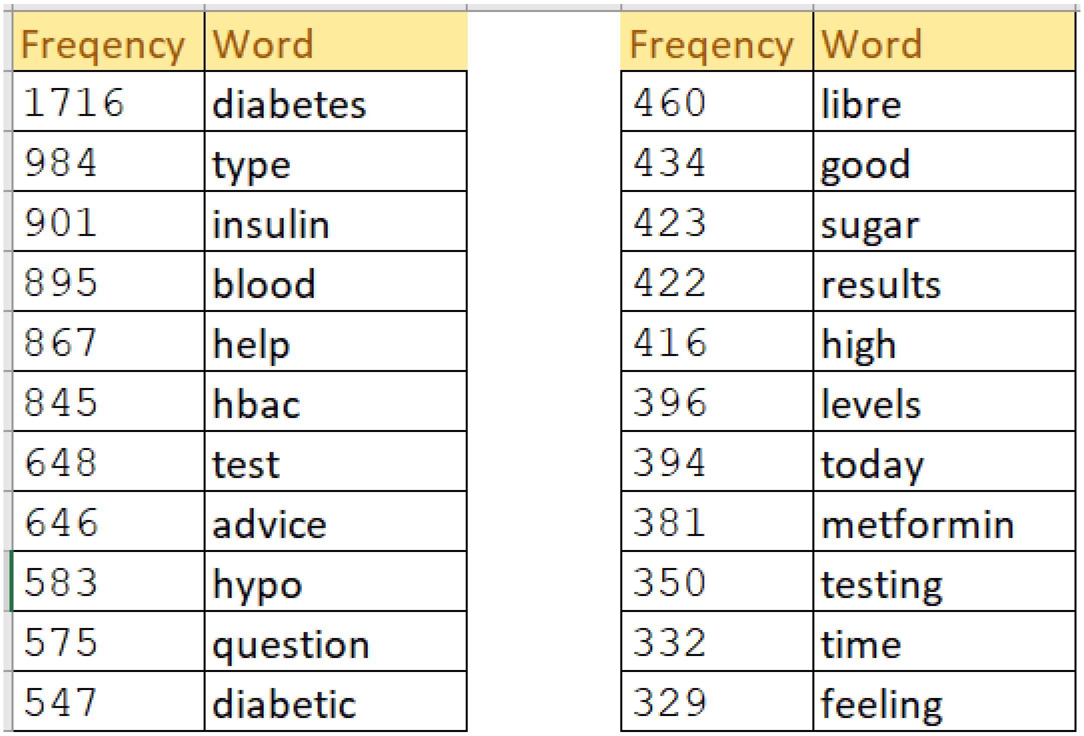


Figure 2.5 Frequency List & Word Cloud Flowchart

We downloaded all the thirty thousand topic headers. In the later sections, we explain in detail the preprocessing. After preprocessing, we have the word frequency table. A partial list of the most frequent words is listed below:



Freqency	Word	Freqency	Word
1716	diabetes	460	libre
984	type	434	good
901	insulin	423	sugar
895	blood	422	results
867	help	416	high
845	hbac	396	levels
648	test	394	today
646	advice	381	metformin
583	hypo	350	testing
575	question	332	time
547	diabetic	329	feeling

Figure 2.6 Partial List of Frequent Terms

2.3.2 DATA PREPARATION AND PREPROCESSING

We have perused through the frequency list up to a frequency of thirty-five or more threads. We could decipher the following five categories, each representing and being identified the presence of multiple words.

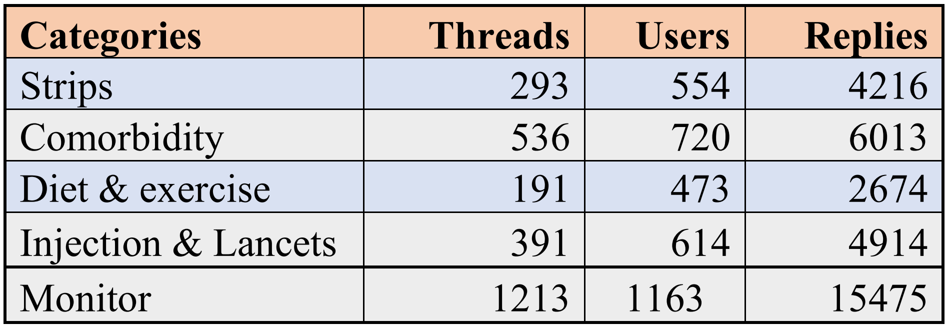
Table 2.1 Categories and the filter criteria

Srl.	Categories	Words
1	Strips	Strip
2	Comorbidity	screening, foot, neuropathy eye retinal
3	Diet & exercise	Food, Carbs, exercise
4	Injection & Lancets	injection, needle, injecting
5	Monitor	libre, meter, accu, monitor

2.3.3 MODEL BUILDING

We have gone through multiple diabetes training videos, books listed in a table elsewhere, and websites to get the domain knowledge. We have filtered the threads based on the presence of the keywords, as detailed in figure 5, into five categories. The number of replies is more than users, as the same user may reply multiple times in the same thread. The Monitors for Glucose testing is the predominant category.

Table 2.2 Category: Number of Threads, Users & Replies



Categories	Threads	Users	Replies
Strips	293	554	4216
Comorbidity	536	720	6013
Diet & exercise	191	473	2674
Injection & Lancets	391	614	4914
Monitor	1213	1163	15475

The data processing is done using python using Beautiful Soup, Selenium, WordCloud, Gensim, and other relevant libraries. All the pages are downloaded and saved as HTML files. The files are named Fnnmm, where nn is the thread ID and mm page number within the thread as displayed on the website.

We developed another module to extract the contents of the posts/replies in the thread. We removed the standard stop words, words of length three or fewer characters, and a few words specific to diabetes. Each thread is considered a document as all the replies are to the same topic answered by different users, and sometimes further clarifications are posted.

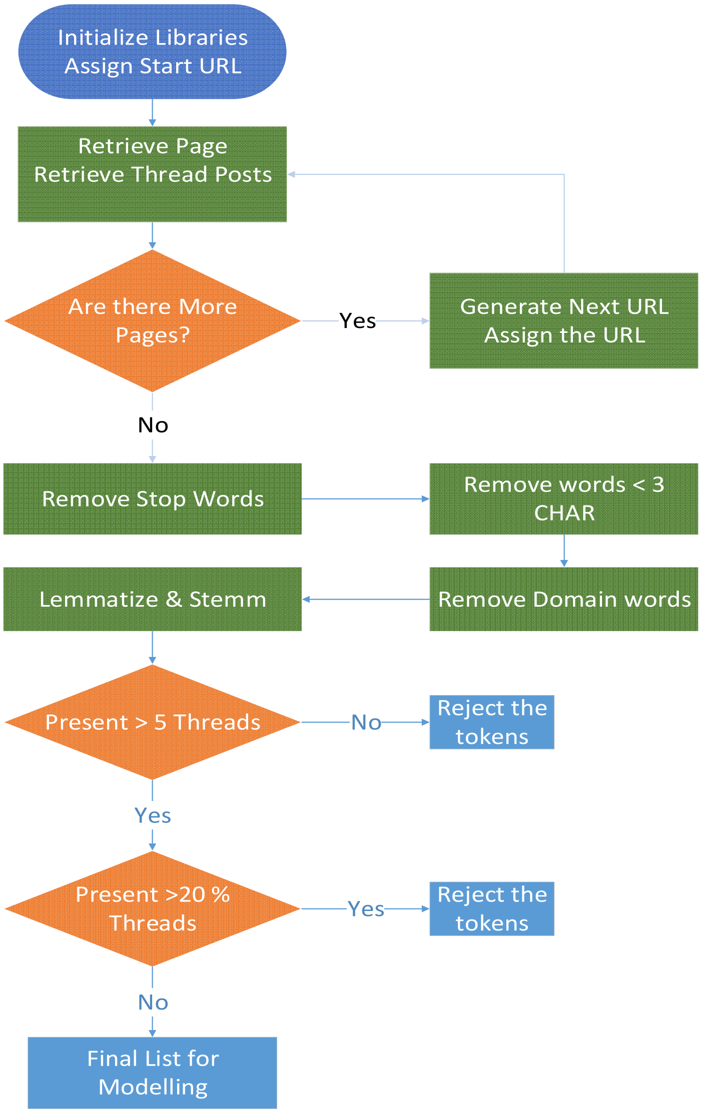


Figure 2.7 Data Preprocessing Flowchart

Lemmatized the words to get the root word to avoid having multiple words with the same root. Then stemmed the words to minimize the number of unique words and further decrease the dimensionality of the data.

For filtrations in the dictionary generation of the corpus, we used the lower limit as a minimum of five documents and a maximum to be present in 20% of documents. If it is present in more than 20%, we filtered them out. We developed two models the first one is Bag of Words (BOW), and the second model is Term Frequency Inverted document frequency (TF_IDF). To comprehend the topics, we have selected ten topics as most of the literature suggests or recommends this number.

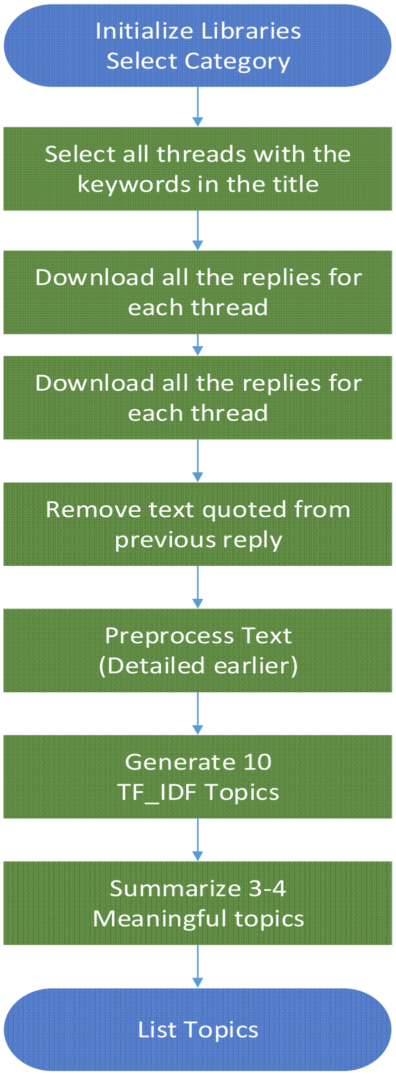
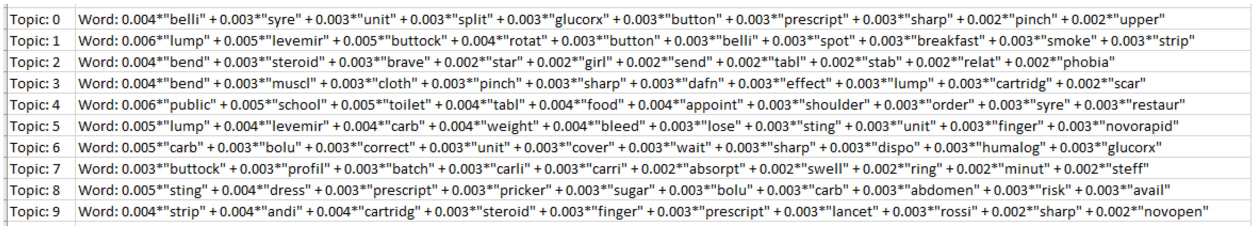

Initialize Libraries
Select Category
Select all threads with the keywords in the title
Download all the replies for each thread
Download all the replies for each thread
Remove text quoted from previous reply
Preprocess Text (Detailed earlier)
Generate 10 TF_IDF Topics
Summarize 3-4 Meaningful topics
List Topics

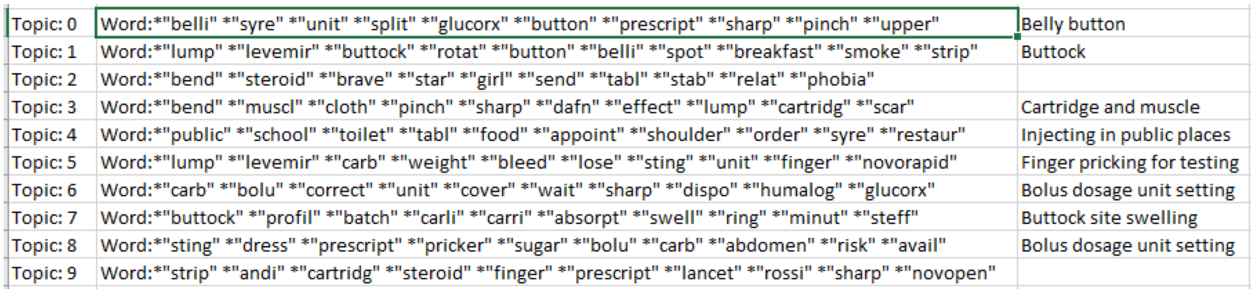
Figure 2.8 Topic Modelling Flowchart



Topic: 0	Word: 0.004*"belli" + 0.003*"syre" + 0.003*"unit" + 0.003*"split" + 0.003*"glucorx" + 0.003*"button" + 0.003*"prescript" + 0.003*"sharp" + 0.002*"pinch" + 0.002*"upper"
Topic: 1	Word: 0.006*"lump" + 0.005*"levemir" + 0.005*"buttock" + 0.004*"rotat" + 0.003*"button" + 0.003*"belli" + 0.003*"spot" + 0.003*"breakfast" + 0.003*"smoke" + 0.003*"strip"
Topic: 2	Word: 0.004*"bend" + 0.003*"steroid" + 0.003*"brave" + 0.002*"star" + 0.002*"girl" + 0.002*"send" + 0.002*"tabl" + 0.002*"stab" + 0.002*"relat" + 0.002*"phobia"
Topic: 3	Word: 0.004*"bend" + 0.003*"muscl" + 0.003*"cloth" + 0.003*"pinch" + 0.003*"sharp" + 0.003*"dafn" + 0.003*"effect" + 0.003*"lump" + 0.003*"cartridg" + 0.002*"scar"
Topic: 4	Word: 0.006*"public" + 0.005*"school" + 0.005*"toilet" + 0.004*"tabl" + 0.004*"food" + 0.004*"appoint" + 0.003*"shoulder" + 0.003*"order" + 0.003*"syre" + 0.003*"restaur"
Topic: 5	Word: 0.005*"lump" + 0.004*"levemir" + 0.004*"carb" + 0.004*"weight" + 0.004*"bleed" + 0.003*"lose" + 0.003*"sting" + 0.003*"unit" + 0.003*"finger" + 0.003*"novorapid"
Topic: 6	Word: 0.005*"carb" + 0.003*"bolu" + 0.003*"correct" + 0.003*"unit" + 0.003*"cover" + 0.003*"wait" + 0.003*"sharp" + 0.003*"dispo" + 0.003*"humalog" + 0.003*"glucorx"
Topic: 7	Word: 0.003*"buttock" + 0.003*"profil" + 0.003*"batch" + 0.003*"carli" + 0.003*"carri" + 0.002*"absorpt" + 0.002*"swell" + 0.002*"ring" + 0.002*"minut" + 0.002*"steff"
Topic: 8	Word: 0.005*"sting" + 0.004*"dress" + 0.003*"prescript" + 0.003*"pricker" + 0.003*"sugar" + 0.003*"bolu" + 0.003*"carb" + 0.003*"abdomen" + 0.003*"risk" + 0.003*"avail"
Topic: 9	Word: 0.004*"strip" + 0.004*"andi" + 0.004*"cartridg" + 0.003*"steroid" + 0.003*"finger" + 0.003*"prescript" + 0.003*"lancet" + 0.003*"rossi" + 0.002*"sharp" + 0.002*"novopen"

Figure 2.9 Output from the TF_IDF for the "Injection and Lancets" category

To get the relevant topics the users are discussing, we removed the probabilities scores of each word from the output and summarized what each topic is about. If the topic is ambiguous, we left it blank. The following is the topic summarization from the word probabilities of "Injection and Lancets."



Topic: 0	Word:*"belli" *"syre" *"unit" *"split" *"glucorx" *"button" *"prescript" *"sharp" *"pinch" *"upper"	Belly button
Topic: 1	Word:*"lump" *"levemir" *"buttock" *"rotat" *"button" *"belli" *"spot" *"breakfast" *"smoke" *"strip"	Buttock
Topic: 2	Word:*"bend" *"steroid" *"brave" *"star" *"girl" *"send" *"tabl" *"stab" *"relat" *"phobia"	
Topic: 3	Word:*"bend" *"muscl" *"cloth" *"pinch" *"sharp" *"dafn" *"effect" *"lump" *"cartridg" *"scar"	Cartridge and muscle
Topic: 4	Word:*"public" *"school" *"toilet" *"tabl" *"food" *"appoint" *"shoulder" *"order" *"syre" *"restaur"	Injecting in public places
Topic: 5	Word:*"lump" *"levemir" *"carb" *"weight" *"bleed" *"lose" *"sting" *"unit" *"finger" *"novorapid"	Finger pricking for testing
Topic: 6	Word:*"carb" *"bolu" *"correct" *"unit" *"cover" *"wait" *"sharp" *"dispo" *"humalog" *"glucorx"	Bolus dosage unit setting
Topic: 7	Word:*"buttock" *"profil" *"batch" *"carli" *"carri" *"absorpt" *"swell" *"ring" *"minut" *"steff"	Buttock site swelling
Topic: 8	Word:*"sting" *"dress" *"prescript" *"pricker" *"sugar" *"bolu" *"carb" *"abdomen" *"risk" *"avail"	Bolus dosage unit setting
Topic: 9	Word:*"strip" *"andi" *"cartridg" *"steroid" *"finger" *"prescript" *"lancet" *"rossi" *"sharp" *"novopen"	

Figure 2.10 Topic summarization from the word probabilities of "Injection and Lancets."

Then we summarized the topics into meaningful groups based on our domain understanding, the below is final result for that of the "Injection & Lancets" category.

Table 2.3 Injection and Lancets Topics

Injection & Lancets
A: Belly button and buttock site injections
B: Injecting in public places
C: Bolus dosage setting and injection

2.4 RESULTS AND CONTRIBUTION

The following are the summarized final topics from the TF_IDF method for all the listed five categories.

Table 2.4 Resultant Topics for all five categories

Categories & Topics
Strips
A: Proper disposal of needles and sharps after the usage
B: The patients are price sensitive and are discussing about pricing online and instore.
C: Discussing how to properly identify faulty batches of strips.
Comorbidity
A: Podiatrist, Footwear and blister or injuries
B: Glasses and blurry vision
C: Neuropathy
Diet & exercise
A: Food for children school going
B: Interaction of different foods and medications
C: Calories and impact of different breakfasts
D: Restrictions on Pizza (Toppings) and Dinner
E: Cholesterol and wine
Injection & Lancets
A: Belly button and buttock site injections
B: Injecting in public places
C: Bolus dosage setting and injection

Monitor
A: Operational issues of the different glucose meters
B: Understand the latest technical features
C: Accuracy and error of these meters
D: Maintenance and spares - Battery, cables, software updates

Considering those topics for “Strips” we found that the patients discussed three significant issues. The American Diabetic Association – Book has discussed these topics as far as strips are concerned. As we can see, the books can be improved by addressing these issues comprehensively. The medical professionals can also address these issues or direct the patients to the relevant, accurate, and authentic sources for questions beyond their expertise.

The users are more concerned with foot and eye problems. Diabetic Patients With Foot comorbidity Fear surgical removal of the foot or lower leg More Than untimely Death (Wukich, Raspovic, & Suder, 2017). The reason may be that the symptoms related to the foot are identifiable by the patients early and are predominant compared to other comorbidities like Kidney, Gum problems & Heart issues. The diabetic books and videos that address the injection site selection and care are outstanding, but the issue of injecting insulin in public places needs to be addressed. As the public generally knows, only drug addicts inject in the open. Like breastfeeding facilities, safe injecting places for diabetes patients should also be provided or encouraged.

Though non-medical but testing is an essential issue for diabetes management. Proper training and guidance by the diabetes educators should be provided, and the responsibility should not be left to the company’s manufacturing the glucose monitors or the sellers of these instruments. Like referring to a dietician, referral to a technician should also be an option, and user guides/manuals in simple, user-friendly language should be made available. There are many books on diabetes-friendly diets. Many dieticians are also diabetic educators and have expert knowledge regarding this topic. The food for Type I diabetic children needs to be addressed more in these books. As far as diet is concerned, more emphasis on the local foods or culture should be stressed. Since it would be very voluminous, an online appendix as a supplement to the books may be developed similar to myfitnesspall application.

2.5 DISCUSSION AND CONCLUSION

The results indicate that the patients need much more education and support than what is currently provided in the existing programs. The analysis shows that they need more help in the non-medical aspects of managing diabetes, such as maintaining the glucose monitors, buying the strips, and disposing of bio-medical waste (Syringes and Lancets). We show that diabetic education could be improved by applying Text Analytics to social media content generated by patients on online social media sites. Clinicians can be trained to address some of these gaps, and for others, they can direct the patients to reliable and expert resources. This framework can be replicated for other chronic diseases to improve the management and assist the patients. We were apprehensive about the user's diabetes education, but the topics discussed show their genuine concerns and significant gaps in the existing literature and programs.

CHAPTER III

ESSAY II: EARLY PREDICTION OF COMORBIDITIES IN CASE OF DIABETIC PATIENTS USING ELECTRONIC HEALTH RECORDS AND ARTIFICIAL NEURAL NETWORKS (ANN).

ABSTRACT

Prevention and early detection of comorbidities in the case of diabetic patients is a primary concern of the physicians as they are more prone to these diseases than the general population. This study applies artificial neural network (ANN) models to the existing medical data stored in EHR to assist medical professionals in predicting and preventing comorbidities. The long-term temporal relations of lab results and medications in electronic health records (EHRs) will enhance accuracy in predicting the progression of related health complications compared to conventional methods that rely on reporting symptoms and clinical diagnostic tests among diabetic patients. Data comprises 17 years of anonymized records from an EHR warehouse having millions of diabetic patients' medical records from the Center for Health Systems Innovation (CHSI) at Oklahoma State University provided by Cerner Corporation. Deep Layered ANN is used to rank the comorbidities using sociodemographic, diagnosis, medication, and lab procedure results over a multi-year observation window of cases. Model performance metrics are compared to regression, random forest, the base machine, and learning models. Our deep-layered ANN models improved comorbidity prediction with an observation window of multiple years.

3.1 INTRODUCTION

Diabetes is a chronic disease whose detection is based on elevated blood sugar. Around 10% of the US population is diagnosed as diabetic, and another 5% is estimated to be diabetic but unaware (Centers for Medicare & Medicaid Services, 2021). It is a chronic condition without a cure, causing additional problems in most cases. Diabetes can be managed by lifestyle changes, proper diet, physical activity, weight management, and medication where necessary. Due to external and environmental conditions and other compulsions, many patients could not make and maintain the necessary lifestyle changes.

Moreover, many patients do not follow the proper diabetic medication regimen as prescribed by their physician. All these actions and inactions lead to comorbidities in the long term. As per Mayo Clinic (2021) the possible complications in the long term for diabetic patients are Cardiovascular (Heart) disease, neuropathy (nerve damage), nephropathy (Kidney disease), retinopathy (eye damage), Lower extremity disease (Foot damage), Skin conditions, Hearing loss, Alzheimer's disease, and Depression.

The management and cure of the above-stated comorbidities are possible and cost-effective if detected in the early stages. Hence, the ADA and the physicians have suggested a screening procedure for diabetes for the detection of comorbidities in the initial stages. Prior research has shown that the patients do not undergo screening as suggested by the physician due to cost, accessibility, and time constraints. There is a significant relationship between regular screening and favorable outcomes in diabetes patients in the case of a screening program for diabetic retinopathy(Zoega et al., 2005).

Diabetic patients regularly visit the primary care provider for their diabetes management and prescription refill orders. Some of the standard tests and screenings that the general practitioner can perform are done during these visits. The screening which requires a visit to an expert or a special lab is the often neglected or delayed.

Bajor and Lasko (2017) have developed a GRU based model to identify omitted medications or billing codes, with the likelihood of such models helping correct them in real-time. Dernoncourt, Lee, Uzuner, and Szolovits (2017) implemented state-of-the-art named-entity recognition (NER) models based on bidirectional long short-term memory (BiLSTM) networks to deidentify clinical notes in EHR so that this resource is available for researchers. Ljubic et al. (2020) developed multiple models to predict the hospitalization of a diabetic patient due to ten comorbidities, the prediction accuracy achieved with the RNN GRU model was between 73% (myocardial infarction) and 83% (chronic ischemic heart disease), while the accuracy of traditional models was between 66% – 76%. They found that the number of hospitalizations was an essential factor for the prediction accuracy, four hospitalizations achieved significantly better accuracy than two hospitalizations.

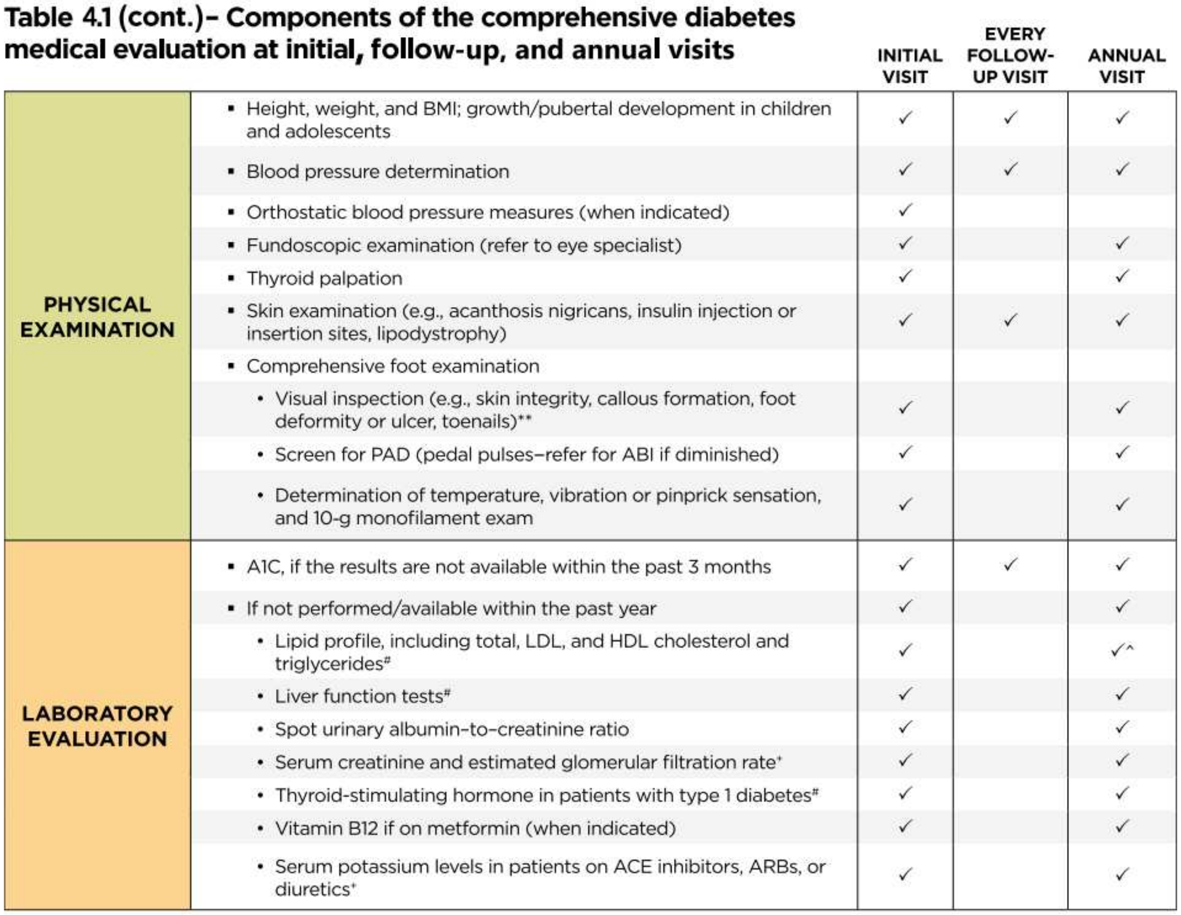


Table 4.1 (cont.)– Components of the comprehensive diabetes medical evaluation at initial, follow-up, and annual visits

		INITIAL VISIT	EVERY FOLLOW-UP VISIT	ANNUAL VISIT
PHYSICAL EXAMINATION	▪ Height, weight, and BMI; growth/pubertal development in children and adolescents	✓	✓	✓
	▪ Blood pressure determination	✓	✓	✓
	▪ Orthostatic blood pressure measures (when indicated)	✓		
	▪ Fundoscopic examination (refer to eye specialist)	✓		✓
	▪ Thyroid palpation	✓		✓
	▪ Skin examination (e.g., acanthosis nigricans, insulin injection or insertion sites, lipodystrophy)	✓	✓	✓
	▪ Comprehensive foot examination			
	• Visual inspection (e.g., skin integrity, callous formation, foot deformity or ulcer, toenails)**	✓		✓
	• Screen for PAD (pedal pulses–refer for ABI if diminished)	✓		✓
	• Determination of temperature, vibration or pinprick sensation, and 10-g monofilament exam	✓		✓
LABORATORY EVALUATION	▪ A1C, if the results are not available within the past 3 months	✓	✓	✓
	▪ If not performed/available within the past year	✓		✓
	• Lipid profile, including total, LDL, and HDL cholesterol and triglycerides#	✓		✓^
	• Liver function tests#	✓		✓
	• Spot urinary albumin-to-creatinine ratio	✓		✓
	• Serum creatinine and estimated glomerular filtration rate+	✓		✓
	• Thyroid-stimulating hormone in patients with type 1 diabetes#	✓		✓
	• Vitamin B12 if on metformin (when indicated)	✓		✓
	• Serum potassium levels in patients on ACE inhibitors, ARBs, or diuretics+	✓		✓

Figure 3.1 Reproduced from (The American Diabetes Association, 2019) standards of Medicare in Diabetes.

Table 3.1 Care Schedule from ADA Complete Guide to diabetes 2nd edition book.

Diabetes Care Schedule	
Every 3 Months	Regular visits to your doctor: If using insulin or if on intensive insulin therapy
Every 6 Months	Glycated Hemoglobin test (HbA1c)
Every Year	HDL/cholesterol: for average reading, more often if high levels are being treated

	Kidneys: microalbumin measured
	Eyes: examined through dilated pupils
	Feet: more often in patients with high risk foot conditions
Every 2 to 3 years	HDL/cholesterol: if last reading indicates very low risk

Reading Turchioe et al. (2020) have identified that visual analogies increase patients' comprehension of changes in their health. Based on this premise, rank-ordering the comorbidities of the specific patient instead of informing the odds ratios or probabilities will make patients comprehend their severity of comorbidity status and motivate them to adhere to the top rank-ordered screenings required in their specific case. Showing further, as this information is specifically tailored for the patient's case instead of graphs and charts in the general population, it might significantly impact the decision-making process for the patient.

3.2 LITERATURE REVIEW

According to the Medical Expenditure Panel Survey, most adults with diabetes have at least one comorbid chronic disease. As many as 40% have at least three (Piette & Kerr, 2006) suggesting that comorbidity is a critical issue in the case of diabetes patients. Earlier research has shown that diabetic patients are twice prone to a heart attack.

The relative impact of glucose levels in the blood (HbA1c) and kidney disease in type 1 diabetes is significantly associated with the length of diabetes.(R. G. Miller, Costacou, & Orchard, 2019) This implies that the duration of diabetes is a predictor of some of the comorbidities in the case of Type I, and we can assume that it may be true in the case of Type II also. Kam and Kim (2017) developed detection models for the early stage of sepsis using deep learning methodologies. They compared the feasibility and performance of the new deep learning methodology with those of the regression method with conventional temporal feature extraction and found their model is better.

Chitravathi and Kanimozhi (2019) have used Sensational Neural Network(SNN) with EHR data to diagnose the most precise disease and if the data is insufficient, SNN would suggest a lab test for additional information to predict. Pham, Tran, Phung, and Venkatesh (2017) have used deep learning to predict the health trajectory and tested their model for diabetes and mental health. Khan and Shamsi (2018) have developed a multi-label classification system of clinical DSS with a natural language process based on family history and EHR data to diagnose not a specific disease only. The data used to train and test the model is only 5000 records, which is very small for training ANN models. Zhang, Chen, Tang, Stewart, and Sun (2017) have studied the related issue of multimorbidity and designed a recurrent decoder to model label dependencies and content-based attention to capturing label instance mapping. This is done to prescribe effective and safe treatment combinations for multimorbidity.

Rasmy et al. (2018) have utilized extensive heterogeneous EHR data to predict the onset of heart failure using RNN-based predictive models. Maragatham and Devi (2019) used an LSTM model to predict Heart failure using EHR data and deep learning models, and they utilized a large dataset of 365K patients. Only 4289 has heart failure. E. Choi, Schuetz, Stewart, and Sun (2017) used RNN for early detection of heart failure using EHR data.

Table 3.2 Neural Networks for predicting diseases

Srl.	Authors & Year	Study	Method
1	Ljubic et al. (2020)	Ten Comorbidities in Diabetes	RNN GRU
2	Kam and Kim (2017)	Early stage of sepsis	Deep Learning
3	Chitravathi and Kanimozhi (2019)	Most precise disease	SNN
4	Khan and Shamsi (2018)	Multi label classification system	NLP
5	Zhang et al. (2017)	Safe treatment combinations for multimorbidity.	Recurrent Decoder

6	Rasmy et al. (2018)	Predict the onset of heart failure	RNN
7	Maragatham and Devi (2019)	Predict Heart failure	LSTM
8	E. Choi et al. (2017)	Detection of Heart failure	RNN
9	Pham et al. (2017)	Healthcare trajectories – Diabetes & mental	LSTM
10	Baumel et al. (2017)	Multi label classification for ICD code	HA GRU
11	Caruana et al. (2015)	Pneumonia Risk and readmission	GA^{2}M
12	Dutta et al. (2020)	Coronary heart disease	CNN
13	Kam and Kim (2017)	Sepsis	Deep Learning

3.2.1 COMORBIDITY IN CASE OF DIABETES

Diabetes causes many other diseases to manifest, mainly if blood sugar levels are not managed over time. Even after proper management, these diseases may manifest, and the probability is very high. The comorbidities considered in this study are Cardiovascular, Peripheral circulatory, Ketoacidosis, Gum Problems, Nephropathy, Ophthalmic, and Neuropathy. As listed earlier, the ADA, CDC, and the physicians recommend regular screening for these comorbidities. American Diabetes Association (2018) has compiled the economic cost of diabetes and came with the detailed cost due to the chronic comorbid conditions.

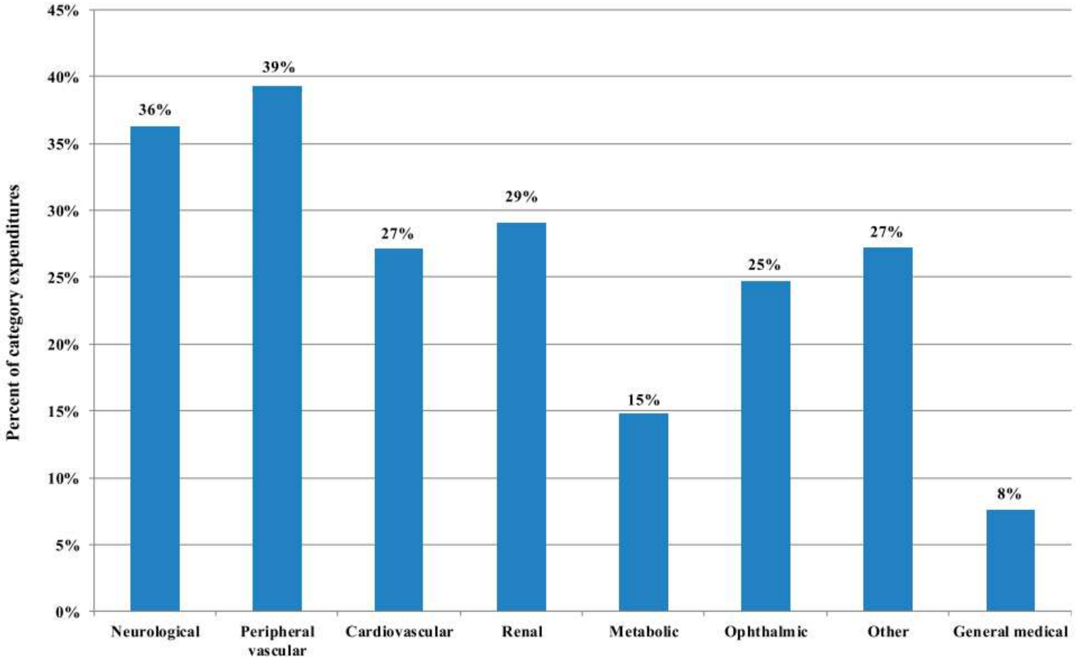


Figure 3.2 Economic costs of comorbid conditions of diabetes reproduced from American Diabetes Association (2018)annual diabetes report

In 2016, 7.8 million diabetes people were hospitalized, 1.7 million with cardiovascular diseases, 130,000 for lower-extremity amputation, and around 37% had chronic kidney disease, 11.7% reported vision disability, including blindness (Centers for Disease Control Prevention & Services, 2020). The conditions and the causes of the studies comorbidities are explained below as per CDC, ADA and other medical organizations.

*Cardiovascular dis*eases (CVD) are a group of disorders of the heart and the major blood vessels. Heart attacks and strokes are usually acute events which and are mainly caused by a blockage in the veins that prevents blood from flowing to the heart or brain. Fatty deposits on the inner walls of the blood vessels that supply blood to the heart or brain is most common reason. Lipid profile test is performed to detect high levels of fatty acids in the blood as a precursor. Strokes may be caused by bleeding from a vein or blood clots in the brain.

Peripheral Circulatory disease is the reduced circulation of blood to the peripheral body parts such as the arms and legs, other than the brain or heart, due to a blocked or narrowed blood vessel. Observed risk factors for this condition include sedentary lifestyle, diabetes, obesity and smoking. Non-clinical intervention includes mostly changes in lifestyle such as achieving normal weight and regular physical activity.

Diabetic ketoacidosis (DKA) is a serious complication of diabetes that can be life-threatening if not detected and treated early. DKA is most common among people with type I diabetes in comparison to type 2. Insufficient insulin to transfer blood sugar into the cells for use as energy causes DKA. Due to insufficient sugar in the cells to use as energy source, liver breaks down fat for fuel, a process that produces acids called ketones. When too many ketones are produced too fast, they can build up to dangerous levels in the body.

Gum (Periodontal) disease is an infection of the tissues that hold your teeth in place. A sticky film of bacteria is formed on the teeth and hardens. Periodontal disease can lead to sore, bleeding gums; painful chewing problems; and even tooth loss in advanced stages. People with diabetes are more likely to have gum diseases, cavities, and other problems with their teeth and gums.

Nephropathy is the deterioration of kidney function of filtering the blood. The final stage of nephropathy is called kidney failure the medical term is end state renal disease (ESRD). Diabetes is the most common cause of ESRD as per CDC. In 2017 more than 288,000 people with ESRD due to diabetes, either had a kidney transplantation or were on chronic renal dialysis. Although type 1 is more likely to lead to ESRD, type 2 diabetes can also lead to diabetic nephropathy.

Ophthalmic diseases, mainly retinopathy and glaucoma, causing vision problems is also a reality for these patients. Neuropathy is the dysfunction or damage of the ending nerves, which causes numbness or weakness in the legs and hands are also faced by people with diabetes, making them unable to do strenuous physical work.

Neuropathy is damage or dysfunction of one or more nerves that typically results in numbness, tingling, muscle weakness, and pain in the affected region. In the case of diabetics, it can cause pain and numbness in the feet to problems with the functions of internal organs.

Diabetic hyperosmolar syndrome is a serious condition caused by high blood sugar levels. Type 2 diabetes patients are more prone to this condition. It is often triggered by illness or infection. In this syndrome, the body tries to decrease the excess blood sugar by passing it through urine. Left untreated, the diabetic hyperosmolar syndrome can lead to life-threatening dehydration. Emergency medical care is essential, and the patient should be taken to ER.

3.2.2 TAXONOMY OF COMORBIDITY STUDIES

Desai, Mehta, Mathias, Menon, and Schubart (2018) as stated that Diabetic Keto Acidosis (DKA) mortality rate has decreased from 0.51%in 2003 to 0.3% in 2014, but the cost has increased from $18,987 to $26,566 per admission in 2014 (after adjusting for inflation).

Balakrishnan, Shakouri, Hoodeh, and Systems (2013) used k-nearest neighbors (KNN) and a decision tree to find the significant risk factors for retinopathy in the case of diabetic patients and found the following variables.

Table 3.3 Critical risk factors in retinopathy predictions as per Balakrishnan et al. (2013)

Variable name	Weight %	Variable name	Weight %
Body mass index (BMI)	80	Low-density lipoprotein (LDL)	38
High-density lipoprotein (HDL)	62	Smoking	23
Triglyceride	60	Alcohol consumption	19
Diabetes duration	60	Alanine aminotransferase (ALT)	10
Glycated hemoglobin (HbA1C)	53	Aspartate aminotransferase (AST)	10
Hypertension	42	Cardiac complication	4
Age	38	Gender	4
Cholesterol	38	Race	4

Piri et al. (2017) have identified many variables and comorbidities as significant factors for retinopathy. In our case, we are rank-ordering probable comorbidities in the case of diabetic patients who do not have any listed comorbidities yet diagnosed. Hence the most critical variables excluding the existing comorbidities are Creatine serum, Blood Urea Nitrogen, and Hematocrit as per this study.

Skevofilakas, Zarkogianni, Karamanos, and Nikita (2010) have studied Type 1 Diabetes Mellitus patients and, based on a hybrid model, have identified the risk factors for developing retinopathy.

Table 3.4 Retinopathy risk factors from Skevofilakas et al. (2010)

Risk Factor	Average ± Standard Deviation
Age	29 ± 12.7
T1DM Duration	8.45 ± 8.07
HBA1C	164.3 ± 60.37
T. Cholesterol	45.77 ± 11.41
Triglycerides	92.45 + 53.04
Hypertension	8% incidence rate
Treatment Duration	0 - 5 years

The U.S. may now be experiencing a regress to the past in the case of lower-extremity amputations, after gaining much success in the past two-decade decline, particularly in men that too the young and middle-aged adults (Geiss et al., 2019).

Song et al. (2019) utilized EHR data to predict Diabetic Kidney Disease (DKD) in diabetic patients.

Table 3.5 Diabetes related comorbidity studies

Srl.	Authors & Year	Comorbidities	Method	Data

1	R. G. Miller et al. (2019)	Kidney and Hb1Ac	Cox Models	DCCT/EDIC Study
2	Piri et al. (2017)	Retinopathy	Variable importance	Cerner Health Facts
3	Skevofilakas et al. (2010)	Retinopathy	Hybrid Model	EURODIAB
4	Rasmy et al. (2018)	Heart Failure	RNN	Cerner Health Facts
5	Balakrishnan et al. (2013)	Retinopathy	Decision tree and KNN	University of Malaya (UM) hospital
6	Song et al. (2019)	Kidney	GBM	HERON

There are multiple studies related to comorbidity in diabetes and other diseases. The studies concentrated on early detection or causes of the comorbidity. Most of the studies have identified correlation but not the causation of the disease. The comorbidities prevalence and the factors contributing to them would help the physicians to give special care and identify at an early stage.

3.3 RESEARCH QUESTION

Rank ordering the probability of the comorbidities for the specific diabetic patient, would assist the doctor in motivating the patient not to miss the higher-ranked screening procedures. This would also encourage the doctor to motivate and refer for the screenings and highlight the necessary lifestyle changes to the patient, which would minimize comorbidity. EHR data has already been extensively used to predict diseases in healthcare.

The EHR data is messy and very high-dimensional. In the third essay, we discussed preprocessing the data for healthcare analytics in detail. For machine learning algorithms to give better results, the data needs to be numerical, non-missing and normalized with low dimensionality. To decrease the dimensionality, we have aggregated the data based on the domain expertise, and the values are collected and generated for this purpose.

3.4 MATERIALS AND METHOD

3.4.1 SELECTION OF DATA

The data for this study is provided by the Center for Health Sciences Innovation (CHSI) at OSU, donated by Cerner. The CERNER Health Facts® database captures and stores longitudinal electronic health record (EHR) patient data. Cerner deidentifies and organizes these data into a data warehouse to facilitate healthcare analysis, research and reporting. This data is aggregated from Cerner and other participating contributing organizations and constitute records from the year 2000 onwards.

CERNER Health Facts® has information on five health outcomes: clinical, economic, process, functional, and satisfaction. This database includes demographics, encounters, diagnoses (ICD codes), prescriptions, procedures, laboratory tests (CPT Codes), locations of services/patients (Clinical department e.g., clinic, ED, ICU, etc.), hospital information (Beds, size) , and billing.

Currently, Cerner Health Facts contains data from:

- Over 65 million patients record
- Patient information from 750 healthcare facilities across the United States
- Over 84 million acute admissions, emergency and ambulatory visits
- Prescriptions of 4,500 drugs by name and brand of 151 million orders
- Over 500 million encounters
- Over 1.3 billion laboratory results
- 4.7 billion laboratory results
- Detailed pharmacy, laboratory, billing and registration data
- 100% of Patients in Orchid, Keck Care and KIDS.

Benefits of using this dataset for research.

- It is a HIPAA-compliant database no need for deidentification.

- A comprehensive source of data generated as a by-product of patient care in real-time hence, eliminating the need for Institutional Review Board (IRB) approval.

- These EHR records are time-stamped and sequenced information useful for longitudinal studies. To deidentify the timestamps of a patient are shifted.

- Helps in determining practice patterns, treatments, and outcomes which is highly suitable for data analytics.

IRB approval for access and use is not required as this research does not involve human subjects and their private information. However, the dataset is de-identified by excluding all the sixteen identifiable values.

The data was queried and filtered based on the following criteria:

1. All the patients who are diagnosed as diabetic, 3.2 million in the warehouse.

2. Patients' who have developed one of the comorbidities being studied, one million.

3. The detection date of diabetes is prior to the detection of comorbidity, 525 thousand.

4. Patients with diabetic medicine or lab tests records are selected for the study.

5. Date range is from 1st January 2000 to 31st December 2017, both ICD 9 & 10 Coding.

6. Data for each patient is till the date of comorbidity detection and not later.

7. The last encounter is used only for the dependent variable or target generation. Medicines and lab tests and other values are not used for independent variable generation.

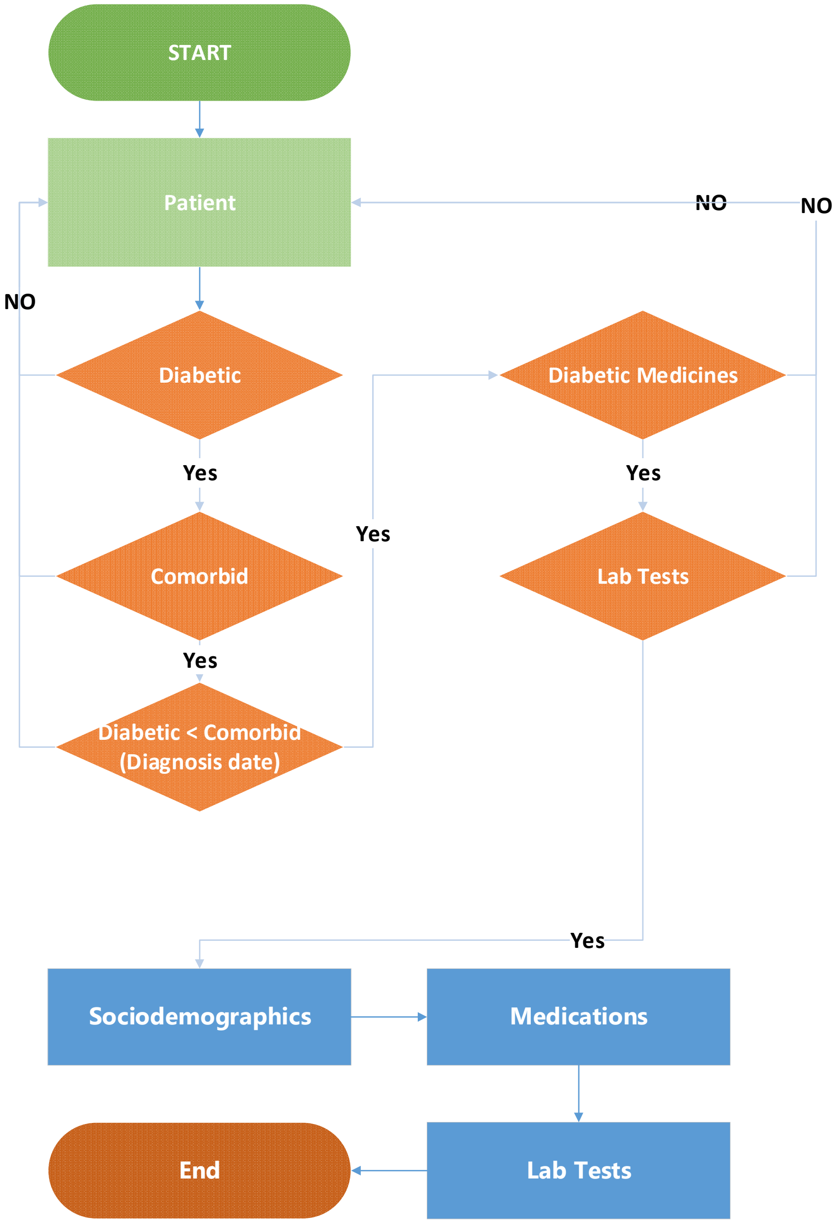


Figure 3.3 Data selection criteria

Table 3.6 Data about comorbidity prevalence in thousands

Srl.	Comorbidity	Patients	Patients with diabetes	Comorbidity detected later	Percentage of the Total
1	Cardiovascular	5,123	787		
2	Peripheral Circulatory	491	332	156	56.3%
3	Ketoacidosis	79	45	12	4.3%
4	Gum Problems	96	10	3	1.1%
5	Nephropathy	183	101	29	10.5%
6	Ophthalmic	109	65	17	6.1%
7	Neuropathy	327	179	56	20.2%
8	Hyperosmolar	23	15	4	1.4%
Total				277	

Cardiovascular disease is the leading comorbid disease among approximately 80% of the patients. Both diabetes and cardiovascular diseases share most of the common factors leading to their sedentary lifestyle, obesity, and improper diet. With this large percentage of comorbidity and shared factors, it is not advisable neglect its screening. Hence, the accuracy of rank-ordering the remaining seven comorbidities and ability to identify among the top two is tested.

As evident from the above table the data is unbalanced and is reflective of the condition in natural population. Stratified over-sampling has been employed to minimize the bias while training the models, since random forest and deep learning models perform well with large data.

3.4.2 MATERIALS

The patient's chronological age is an important predictor of the comorbidities as it signifies the aging of all the organs and systems of the human being. The diabetic age is also essential as it signifies the duration of abnormal sugar levels in the body. Many medical studies have shown that "race" is a significant predictor or contributor to disease due to genetic makeup or the cultural conditions of food intake and other practices. Gender is also a significant variable in disease vulnerability and differentiation.

Many medical studies have found marital status as a significant factor, and it has been shown that cohabiting persons are less prone to health issues and recover faster after falling sick. Adherence to medical regimens is also very high. Rural residents generally have less access to good medical facilities, and due to the distance and time necessary to reach a hospital, they delay medical attention. Prior studies have shown that teaching hospitals have better outcomes as they are more into research and are probably less profit-motivated. Hence patients visiting a teaching hospital would indicate that their case is more complicated or is willing to go that extra mile to achieve better health outcomes.

Table 3.7 Demographic profile of the cohort

Gender		Rural / Urban		Race		Insurance	
Male	47.53%	Rural	20.07%	African American	19.32%	Government	51.85%
Female	52.47%	Urban	79.93%	Caucasian	67.15%	Private	42.17%
				Other	13.52%	Self	5.98%

Obesity, as measured in body mass index (BMI) units, is also known to be a general indicator of health, and diabetes and obesity have common causal agents. Obesity is directly linked to the economic income of the countries. Developing countries like China and India have higher rates of obesity(Levine, 2011). Excess alcohol consumption has been associated with higher mortality and other diseases. Obesity is a significant predictor of Type II diabetes and cardiovascular diseases. Hence, we are considering them also into account in our study.

Table 3.8 Lifestyle composition of the cohort

	Smoker	Alcoholic	Obese
Yes	16.84%	5.89%	13.53%
No	83.16%	94.11%	86.47%

Prior treatment, dosage strength and duration are also general indicators of the progression of diabetes and the person's overall health. So, to assess the treatment plan, we have categorized the diabetes medicines into different categories. The categories along with the effective medications (molecular) under these categories are detailed in the table 3.4.

Table 3.9 Categorization of Diabetic Medicines

Category	Medicines (Generic name)
Alpha-glucosidase inhibitors	acarbose, miglitol
Amylinomimetic	Pramlintide
Biguanides	Metformin
Dipeptidyl peptidase-4 (DPP-4) inhibitors	alogliptin, linagliptin, saxagliptin, sitagliptin
Dopamine agonist	Bromocriptine
Glucagon-like peptide-1 receptor agonists	albiglutide, dulaglutide, exenatide, liraglutide, semaglutide
Insulin	Insulin
Meglitinides	nateglinide, repaglinide
Sodium-glucose transporter (SGLT) 2 inhibitors	dapagliflozin, canagliflozin, empagliflozin, ertugliflozin
Sulfonylureas	glimepiride, gliclazide, glyburide, chlorpropamide, tolazamide, tolbutamide
Thiazolidinediones	rosiglitazone, pioglitazone

Many lab tests are performed regularly for diabetic patients to monitor their blood sugar levels and other important factors. The values of the critical indicators show the progression of diabetes and some the comorbidity. The following is the list of lab tests and what they measure for which comorbidity.

Table 3.10 Lab tests performed for Diabetes and Comorbidity screening

Srl.	Comorbidity	Lab Test
0	Diabetes	A1C
		Fasting Blood Sugar
		Glucose Tolerance
		creatinine blood
1	Cardiovascular	Lipid Profile
		ECG
		Treadmill Tests
2	Peripheral Circulatory	Ankle-Brachial Index test (ABI)
3	Ketoacidosis	Spot urinary albumin-to-creatine ratio
		Serum creatine and estimated glomerular filtration rate
4	Gum Problems	None
5	Nephropathy	Liver function
6	Ophthalmic	Fundoscopic
7	Peripheral Neuropathy	Nerve Conduction Velocity and Electromyography

The variables used in this experiment are sociodemographic, medicines, and lab tests. The sociodemographic variables are not readily available in the EHR systems. The aggregation of medicines is complex as the generic names and combinations make the dimensionality reduction difficult. The lab tests are complicated as in each test, there are many values reported in numerical, and categorically as in the case of A1C, the number is in percentage and categorical (Diabetic, Prediabetic, or Normal). The medical literature highlights the importance of all the reported values. It is not easy to summarize the test results and get one numerical or a categorical value to represent the test. We summarized the number of times the tests were performed as normally the frequency of testing is low if prior tests were normal.

3.4.3 METHOD

Logistic regression, Decision Tree, Random Forest, KNN, and simple NN were tested for the prediction. Random Forest was able to better classify among all these standard multiclass classifiers. We compared the random forest with the deep layered neural network predictor. The ANN has four hidden layers, with 40, 30, 20, and 10 neurons at the respective layers. The last layer has a SoftMax function with neurons per the number of classes. Increasing the layers or neurons did not affect the accuracy significantly.

Data storage, retrieval, filtration and minor transformations were in relational database using structured query language (SQL). The analysis is done in python version 3.7.3 using the modules scikit-learn version 0.22.1 for splitting the data and to run the K cross fold, machine learning, imblearn 0.0 for oversampling, and keras 2.3.1 for ANN modeling.

Our data had sociodemographic, Medicines and Lab tests. We ran three different models to assess the change in accuracy along with the system's complexity. Model A has variables from all three categories, Model B has sociodemographic and medicines, and Model C has sociodemographic and lab tests only. For both random forest and ANN, the model performed better.

The probable major comorbidities can be regrouped based on the specialist or facilities needed to perform the screening. We reclassified the target and reduced to five only. We performed the analysis on both the comorbidities and screening categories.

3.5 RESULTS

The ANN model with all the three category of variables a better overall accuracy in predicting the screening requirement. The 67.7% may seem not good overall but the detailed analysis shows its utility to the healthcare community.

Table 3.11 Accuracy of the Random Forest and ANN models

		Comorbidity		Screening	
Model	**Categories**	**Random Forest**	**ANN**	**Random Forest**	**ANN**
A	Sociodemographic, Medicines, and Lab Tests	51.1%	58.5%	59.3%	67.7%
B	Sociodemographic, and Medicines	49.0%	53.6%	60.0%	63.0%
C	Sociodemographic, and Lab Tests	50.3%	53.9%	58.6%	64.1%

The prediction of probable comorbidity overall is at low 59.3% for the ANN model. It is very good at predicting Neurological, Nephropathy and Ophthalmologist comorbidities. Let's consider the case of Ophthalmologist (mainly retinopathy), this system is able identify 43% accurately while only 5.1% are part of the population.

Table 3.12 Comorbidity prediction accuracy for Random Forest and ANN

Comorbidity	Population	Random Forest	ANN
Peripheral Circulatory	60.1%	56.2%	64.0%
Ketoacidosis	3.9%	45.9%	29.7%
Gum	1.0%	44.4%	9.7%
Nephropathy	9.3%	61.2%	56.5%
Ophthalmologist	5.1%	29.6%	43.8%
Neurological	19.5%	37.5%	54.1%
Hyperosmolarity	1.1%	30.9%	1.2%
Overall		51.1%	58.5%

The prediction of probable screening overall is at low 69.4% for the ANN model. It is very good at predicting Neurologist but not so good at Ophthalmological comorbidities. This model is able to highly predict where the ABI Test is needed.

Table 3.13 Screening prediction accuracy for Random Forest and ANN

Screening	Population	Random Forest	ANN
ABI Test	60.1%	61.4%	71.5%
Blood & Urine Tests	14.3%	54.1%	65.8%
Dentist	1.0%	59.6%	6.7%
Neurologist	19.5%	58.7%	77.9%
Ophthalmologist	5.1%	46.1%	14.9%
Overall		59.3%	69.4%

3.6 EXPLAINING THE DEEP LEARNING MODEL

The deep learning models are black-box models as we do not know or can easily comprehend the algorithm and explain how the prediction is obtained. This complex nature and inability to explain causes the low adoption of deep learning models, specifically in situations where the errors would be very costly. Based on the specific context, the costliness of Type I or Type II error depends on medicine and healthcare. Hence most of the algorithms are developed as Decision Support Systems (DSS), and the final decision is left to the human expert, in this case, the physician.

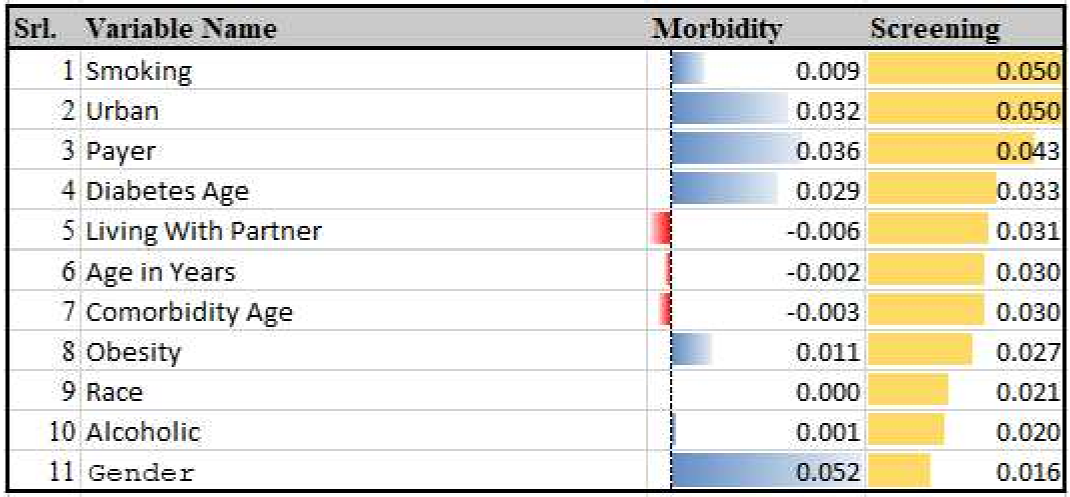
It would be helpful to the human expert to understand the prediction if we can explain how the model is making the prediction. There is extensive research on explainable Artificial intelligence (XAI).T. J. A. I. Miller (2019) has argued that XAI should not be from the researcher's perspective. It should follow the social sciences, which have already researched how people define, generate, select, evaluate, and present explanations. He argues that people employ certain cognitive biases and social expectations in the explanation process. Barredo Arrieta et al.

(2020) have summarized the present Responsible Artificial Intelligence literature, namely, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability, and accountability at its core.

XAI has two methods of explanation one is local the other is global. The local explanation is to explain the factors based on which the model has given the prediction in the specific case (sample). The global methods explain the contributing factors and their weights for the model's prediction. One of the methods is calculating the variable importance based on one left out. In this method, the model's accuracy is calculated by leaving out one of the variables, the decrease in accuracy is attributed to the contributing factor of that specific variable.

Ancona, Ceolini, Öztireli, and Gross (2019) discuss the theoretical properties of several attribution methods and show how they share the same idea of using the gradient information as a defining factor for the functioning of a model, and show the strengths and limitations of these methods and compare them with available alternatives. Caruana et al. (2015) have shown that it’s possible to be accurate and explainable with the machine learning models using high-performance generalized additive models with pairwise interactions (GA2Ms) in real healthcare problems. We have implemented this method, and the following is the variable importance chart for the Model A which is most accurate in rank ordering the comorbidities.

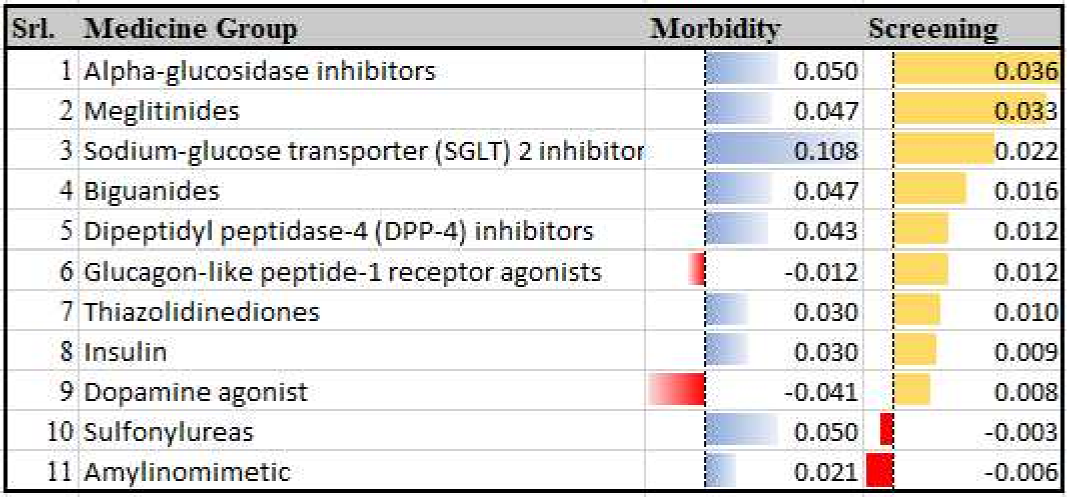
Table 3.14 Variable importance (Sociodemographic) for Model A



Srl.	Variable Name	Morbidity	Screening
1	Smoking	0.009	0.050
2	Urban	0.032	0.050
3	Payer	0.036	0.043
4	Diabetes Age	0.029	0.033
5	Living With Partner	-0.006	0.031
6	Age in Years	-0.002	0.030
7	Comorbidity Age	-0.003	0.030
8	Obesity	0.011	0.027
9	Race	0.000	0.021
10	Alcoholic	0.001	0.020
11	Gender	0.052	0.016

Based on the above chart, we can see that smoking, race, location, and obesity are important variables hence they should be recorded with accuracy in the EHR for proper prediction. Further, a detailed analysis should be done to assess the importance of further sub-divisions.

Table 3.15 Variable importance (Medicine groups) for Model A



Srl.	Medicine Group	Morbidity	Screening
1	Alpha-glucosidase inhibitors	0.050	0.036
2	Meglitinides	0.047	0.033
3	Sodium-glucose transporter (SGLT) 2 inhibitor	0.108	0.022
4	Biguanides	0.047	0.016
5	Dipeptidyl peptidase-4 (DPP-4) inhibitors	0.043	0.012
6	Glucagon-like peptide-1 receptor agonists	-0.012	0.012
7	Thiazolidinediones	0.030	0.010
8	Insulin	0.030	0.009
9	Dopamine agonist	-0.041	0.008
10	Sulfonylureas	0.050	-0.003
11	Amylinomimetic	0.021	-0.006

3.7 DISCUSSION

In this study, we proposed a new approach to prioritize screening in the case of diabetic comorbidities by rank-ordering as prioritizing the top two. The rank-ordering is done by ANN utilizing the existing sociodemographic, medical prescriptions, and lab tests.

While predicting the comorbidities and significant factors leading to them was studied earlier for individual diseases, the present is the first one that employs ANN and EHR data to prioritize screening. The results we obtain suggest that the combination of EHR data and ANN models better prevent diabetic retinopathy, nephropathy, and neurological problems.

Singh and Varshney (2020) in their systematic review of IT-based reminders for medication adherence, have suggested that this can be utilized for screening adherence, too, with reminders being explicitly auto-generated for the patient using the EHR system. Reminding patients of the missed screenings based on EHR data increased the adherence rate by 20% in the case of breast cancer patients (Jain, Guan, FaisalUddin, Manoucheri, & Fang, 2019). We can attain better screening compliance by implementing a similar reminder system for the top rank ordered comorbidities.

This framework of developing a predictive model using EHR data in health care can be generalized to other chronic conditions such as dementia and Alzheimer's. With the increase in the longevity of the population and higher share of seniors, this framework may be used for the predictions of geriatric conditions.

3.8 LIMITATIONS

The results would be more accurate if we had the patient's weight management and physical activity information. Since it is HIPPA compliant, we do not have location information.

Past studies have shown that a person's location does show the probability of the disease. Highly industrialized/manufacturing areas cause more respiratory infections.

The date range is from 1st January 2000 to 31st December 2013, so we have only the ICD 9 Coding. More data can be utilized by merging with ICD 10 coding beyond 31st December 2013. A robust ICD 9 AND ICD 10 diagnosis code matching algorithm or procedure must be implemented.

Piri, Delen, and Liu (2018) has developed a better algorithm to address the imbalanced data problem by developing the Synthetic Informative Minority Over-sampling (SIMO) algorithm. This algorithm is very computing resource-intensive and useful when the numeric values are continuous. Since most of our variables were not real numbers by design, we could not use the algorithm. In future studies, we will incorporate this algorithm for oversampling.

Foot examinations and other such physical examinations which are not performed by Lab tests and results may not be available in lab procedures are not included in this model. These examination results are usually entered as clinical notes in the EHR, and they have much vital information prior to ordering for actual tests. Text Analytics could extract information from clinical notes on such examinations' performance and, if possible, on the result. These variables would enhance the prediction model. As Van Calster, Wynants, Timmerman, Steyerberg, and Collins (2019) stated in their paper, we have provided all the methodology for any other researcher to replicate or improve our DSS. As the data is proprietary to the Cerner corp., we are not able to share its contents.

CHAPTER IV

ESSAY III: BEST PRACTICES TO DEAL WITH THE CHALLENGES OF DATA PREPROCESSING FOR ANALYTICS STUDIES IN HEALTHCARE

ABSTRACT

Data preprocessing is an essential step for machine learning algorithms in data analytics. The selection of the preprocessing methodology is based on the data collection methodologies and the models that are employed in the study. The standard practices of preprocessing for machine learning may not work effectively in the case of healthcare data as the data collection and interpretation of the results is quite different from that of general business problems. Human life is invaluable and its worth cannot be estimated or valued in terms of dollars. Misinterpretation in the case of healthcare decisions is loss of human life or major injuries and should be avoided at all costs and levels. This study explores the standard preprocessing methodologies and their impact in the case of healthcare data analytics. The current study proposes best practices based on past research and the experience obtained in the projects discussed earlier in the thesis.

4.1 INTRODUCTION

Data in the real world is dirty and corrupted with inconsistencies, noisy, incomplete information, and nonstandard and contains missing values. It may be aggregated from diversified sources (in prior study, the data is from a single application but different locations) using data mining and warehousing techniques. Preprocessing the noisy data is essential to obtain quality results and identify patterns effectively. This is a necessary step in Machine Learning as the quality of data and the valuable information that can be derived from it directly affects the ability of the model to learn and predict with accuracy. Therefore, it is essential that we preprocess our data before training and validating the model in an appropriate method.

Idri, Benhar, Fernández-Alemán, and Kadi (2018) have performed a systematic map of medical data preprocessing for data mining methodology papers. They have summarized the findings from 110 studies. There are similar other systematic review studies. However, they have not highlighted and contrasted the standard machine learning practices and the modifications done or needed in case of health data.

The HIPPA regulations add another layer of data preprocessing to the medical data. The HIPAA Privacy Rule was established to protect individuals' medical records and other individually identifiable health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically. Healthcare Analytics use health data, and these are also subject to this regulation.

If HIPPA compliant de-identified data is available than it resolves most of the issues with deidentification and IRB approval process. Otherwise the 18 identifiers that make health information PHI: Names, Dates- except year, Telephone numbers, Geographic data, FAX numbers, Social Security numbers, Email addresses, Medical record numbers, Account numbers, Health plan beneficiary numbers, Certificate/license numbers, Vehicle identifiers and serial numbers including license plates, Web URLs, Device identifiers and serial numbers, Internet protocol addresses, Full face photos, and similar images, Biometric identifiers (i.e., retinal scan, fingerprints, DNA Sequence, Dental impressions), Any unique identifying number or code must be removed (Alder, 2022).

Absence of some of the PHI variables is a limitation. Specifically, the absence of Geographic data and device information is limiting. The absence of geographic data limits us from studying county, state, rural/urban, and other disparities. It also limits us from studying the accessibility and effect of location or other services on healthcare. Many machine learning algorithms for De-identification of clinical notes have been studied but are not adopted widely by the medical community. Hence, we have a paucity of good repository of Clinical notes for research.

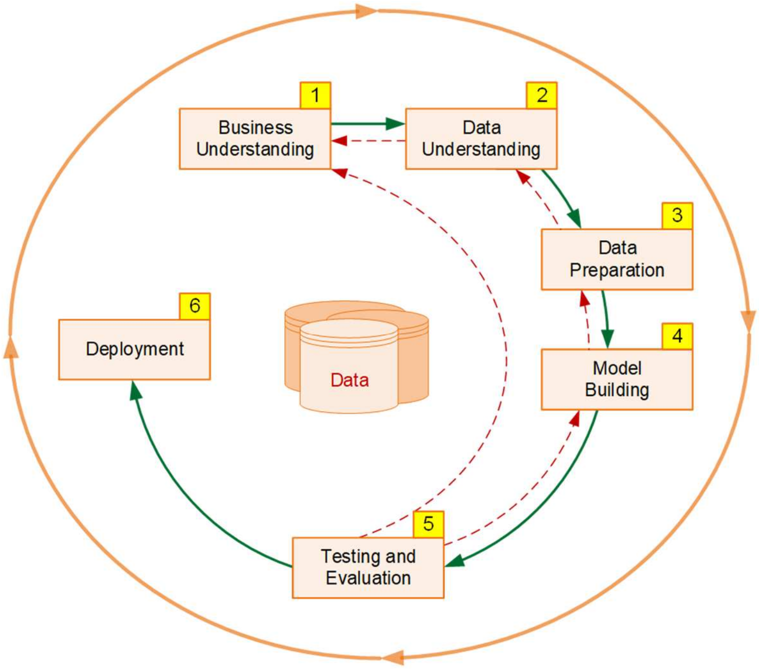


Figure 4.1 Cross-Industry Standard Process-for Data Mining (Delen, 2021)

Healthcare Analytics can be considered a subdomain of data mining. The CRISP-DM process is an industry verified and tested methodology followed by many professionals for data mining. Healthcare Analytic researchers and project developers can streamline their work by following this flow.

The first two steps of business understanding and data and understanding are crucial. Healthcare is an expert professional rendered service. It is different from standard businesses. Seventy percent of the hospitals in the USA or not for profit, so their main objective is not profit maximization. Hence, we need to understand the goals and limitations of these organizations. As a critical and lifesaving sector, this industry has multiple regulations that should be understood and adhered to while collecting and analyzing the data.

Data preprocessing best practices discussed in this chapter cover the steps data understanding and data preparation under the CRISP-DM. The first step under data understanding is to describe the data in terms of quantity (Rows or observations), Value types (numeric, categorical, Boolean, date), Coding schemes of gender, payer. ICD coding and others. The second sub task is to generate tables and charts to know the distribution and availability of different segments, this would guide us the need of dimensionality reduction.

4.2 BACKGROUND

Many studies focused on data preprocessing and suggested best practices. Goldstein, Navar, Pencina, and Ioannidis (2016) have done a systematic review on opportunities and challenges in developing risk prediction models with electronic health records data. Payne et al. (2015) have summarized that with the broad adoption of EHR, clinicians are voicing concerns about unintended clinical consequences, reduced time for patient-clinician interaction, and lengthened clinician workdays while the promised interoperability between different EHR systems is not achieved. Hence, while using the EHR data, care should be taken that the data might not be of the gold standard because of the concerns mentioned above and others.

Fan, Chen, Wang, Wang, and Huang (2021) have comprehensively reviewed data preprocessing techniques for analyzing massive building operational data, similar to what we are doing for healthcare analytics. They have identified comprehensive preprocessing techniques in terms of their applications in missing value imputation, outlier detection, data reduction, data scaling, data transformation, and data partitioning.

4.3 DATA PREPROCESSING

Preprocessing or data preparation, among others, includes handling Null Values, Standardization, Categorical Variables, One-Hot Encoding, Multicollinearity, missing values, normalization duplicates, and Outliers. The data preprocessing can be subdivided into six categories:

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Dimensionality Reduction
5. Longitudinal variables creation
6. Textual Data handling

Table 4.1 Data preprocessing steps commonly done

Data Cleaning	**Data Integration**	**Data Transformation**
Missing Values	Multiple Sources	Generalization
Outliers	Different Periods	Normalization
Noisy	Different Units	Aggregation
		Attribute Selection

Dimensionality Reduction	**Longitudinal Variables**	**Textual Data**
PCA	First/Last	Spell Check
Discretization	Mean/Median	Abbreviation Expansion
Attribute subset	Min/Max	Stemming
Binning		Lemmatization

4.3.1 DATA CLEANING

The data cleaning involves handling Missing Values, Noisy Data, and Removing Outliers.
It is necessary to filter out irrelevant or erroneous records, and sometimes it may be due to a simple data entry mistake. In a recent project related to Polycystic ovarian syndrome (PCOS), few of the patient's records were marked "male". It is certain that PCOS could not happen in a male and the probability of error in diagnosis is very low in comparison to the gender selection while data entry hence we marked all the records as "Female" and did not remove the records marked as "Male".

The need to filter records based on age or other criteria for the research question being addresses may arise. In the case of infertility treatments, we select the patients within reproductive age group of 16 to 45 years, similar selection should be made based on condition being studied. It is customary to record sociodemographic values during the first visit at every establishment in healthcare settings. This may be done due to the unavailability of the recorded data from previous organizations or to record changes if any. Due to time constraints and frequent visits, all the sociodemographic values are not captured on each visit.

Data cleaning also implies selecting the best value for a variable when we have multiple probable values. At times the value of a variable may vary genuinely and at other times they may have been entered as per the user's perception, or the user gives varying answers. Some of the important demographic information is not captured at all due to non-relevance to healthcare needs such as Income and Education. Many studies have found the significance of income and education level on health outcomes.

Noisy data is meaningless data that cannot be understood and interpreted correctly by algorithms. Normally healthcare data is not noisy as it is collected by medically trained personnel. This may arise when machine inputs are stored in EHR, such as temperature and other clinical information, and the machines are not functioning correctly. Similarly, software transcribed medical notes may be inaccurate and gibberish. Data collection and recording methodologies should be recorded and evaluated to avoid noisy data.

Race of a person does not change over time, but we have seen different races being recorded for the same person in EHR data. It implies the data recording was wrong, the person is of mixed race, and the perception-based entry was used. The most frequent value may be used, or another category of others "mixed race" should be assigned to such patients. Celis, Keswani, and Vishnoi (2020) detailed how to mitigate bias by data preprocessing. In AI, it is essential to minimize race and gender bias in case of hiring and criminal justice systems. However, in healthcare, we do not want to remove the racial bias from diagnosis as it is biological or natural and would help us in proper treatment and prediction. A specific example is that it has been found that South Asians are prone to heart problems having the BMI and waist ratio measurements being considered normal for the US population.

Residency of the patient in Urban or Rural areas has been found to impact the health outcomes due to access, pollution or other reasons. However due to HIPPA and other regulations, this information is generally not made available from EHR data to the researchers. The hospital location can be used as a proxy for patient residence, even though a rural patient would visit an urban hospital, and the reverse is not so often true. Broadband access is also limited or spotty in rural areas, which reduces the ability to utilize the latest technological advances. Summers-Gabr and Policy (2020) , during their COVID19 mental health disparities study, have found that the 20 million people who do not have broadband access faced a different set of barriers, including access to telehealth and other facilities.

Bias is not limited to race only, in case of healthcare the rural population is underserved and has accessibility issues too. Seker, Talburt, Greer, and Informatics (2022) found that the diagnosis of multimorbidity among patients decreased with rurality. The diagnosis is missing but we do not have methods to rectify hence while training and predicting with rural data care should be taken. One option is to train the models with urban data as the accuracy would be high,

Obesity is another socio-demographic variable that is difficult to capture and have many issues. Several medical conditions are impacted by obesity, we have ICD coding for BMI, obesity and morbidly obesity. It is coded in terms of BMI or obesity level and the ICD coding indicates its presence, but absence of coding does not imply the reverse, that is the patient is normal. Martin, Chen, Graham, and Quan (2014) have found that when coded the obesity is accurate in EHR data, but it is not coded in all the cases when the patient is obese. The physician may not measure and enter obesity information if the condition is not impacting the treatment or necessary for the insurance claims. So, based on the question and in how many cases the BMI or obesity is recorded, the decision to use this information should be considered.

Smoking and alcoholism information is not coded fully. In case of smoking normally it is coded as smoker, we do not have the information about the start and the number of packs being smoked in a day. Similarly, for alcoholism also we do not have the consumption start and at what frequency and how much quantity of alcohol is consumed. Smoking and Alcoholism are coded if they are severe and impact the diagnosis or treatment, and sometimes they may be recorded in the clinical notes, which are rarely made available for researchers along with the EHR data. McVeigh et al. (2016) assessed the validity of obesity, smoking, depression, and influenza vaccination indicators from an EHR surveillance system and found the results to be significantly accurate. This finding is contradictory to our experience in these projects, hence usage of these variables should be validated.

Over the counter (OTC) drugs are normally recorded in the active medication list, but their dosage and duration are not recorded. Physical activity level and types of exercise and its duration, and intensity is also missing in most EHR data. It may be present in the clinical records which are mostly not available due to HIPPA compliancy. This information if available should be validated for accuracy.

Technical jargon is prevalent in Healthcare like any other scientific field, healthcare has its own vocabulary and abbreviations. Domain knowledge is essential to understand the healthcare notes and data. In addition, each specialty has its own set of additional abbreviations and jargon. This is a significant impediment in healthcare analytics. The language and technical jargon are daunting, and the learning curve is steep. The domain knowledge is vital to doing research in any field, and it is much more in healthcare. Before doing the research on diabetes, going through the listed books multiple times to comprehend the technical nuances of the field assisted very much.

Table 4.2 List of Diabetes books by experts

Srl.	Title	Publishing Year
1	The Medical Library Association Guide to Finding Out About Diabetes: The Best Print and Electronic Resources (Guide to Finding Out About Diabetes)	(2013)
2	The uncomplicated guide to diabetes complications	(1998)
3	Diabetes research: a guide for postgraduates	(2000)
4	Complete guide to diabetes: the ultimate home diabetes reference	(1999)
5	Mayo Clinic: the essential diabetes book: [how to prevent, control and live well with diabetes]	(2009)
6	Obesity and type 2 diabetes mellitus	(2012)
7	Diabetes A to Z: what you need to know about diabetes, simply put	(2010)

The best way to understand the technical jargon is to attend an introductory course on the subject, read multiple books on the topic and consult a subject matter expert on important technical issues and if possible, include them in your project. Pitt and Hendrickson (2021) have clearly stated the reasons for having the medical jargon in notes along with the explanations, hence proper care should be dealt to handle the jargon and the explanations.

Missing Values is a significant problem in data analytics, and the missingness can be classified into three categories Missing completely at random, missing at random, and Missing, not at random. Missing completely random implies that the data sample is likely representative of the population, and we can run the analysis without bias by dropping these observations. Missing at random is the state that there is complete information when variables can fully account for the missingness. This assumption cannot be statistically or by other methods verified and is presumed based on observations and rational reasoning. Data collection should be improved to avoid missingness in these cases. Missing not at random (MNAR) is all other cases. Samuelson, Spirer, and straight (1992) discussed how missing and/or distorted data about health, demographics, and law enforcement could indicate patterns of human rights violations. So MNAR cases in healthcare should be avoided to get unbiased results.

Multiple methods have been developed and suggested to handle missing data based on the reasons for missingness. We should first understand the data, how it is collected, and the possible reason for the missingness. Then we should handle it accordingly, and also, we should understand how our model/algorithm would handle the missingness if left untreated. Regression and deep learning models cannot use observation with missing data and they will drop those observations. So, either we have to discard these observations or impute them. Before imputing them, we need to understand how the missingness is occurring random, specific situations, data not recorded, or value not present.

Imputation can be done by regression, random forest, or other standard imputation methods. In healthcare metrices we have nationally reported figures by CDC. Instead of replacing the values from our sample we should use the nationally representative reported facts. In case of obesity, CDC estimates 40% are obese and severe obesity to 10%. Hence, we should impute probability of obesity first. Them the top 40% probabilities should be assigned 1 and others 0 for the Obesity dummy variable. By adopting this approach, we are keeping in line with the population metrics.

It has been observed that the fact that the value is missing is very significant hence a dummy variable indicating the missingness is also useful and should be utilized in the analysis. To address the issue of missing values in EHR data, Piri (2020) has suggested a framework named '*Missing Care*' based on their experimentation on EHR data in the case of Parkinson's disease, which is a chronic disease similar to diabetes.Mirkes, Coats, Levesley, and Gorban (2016) have developed non-stationary Markov models to address missing data completely at random and highlighted its efficacy on Trauma Audit and Research Network (TARN) database, the largest trauma database in Europe.

4.3.2 DATA INTEGRATION

This step involves integrating data from multiple sources, locations, or tables. Li and Ngom (2015) have delved into data integration of in case of healthcare and machine learning in detail. In our study, the data is integration is done by the Cerner Corp., they have merged data from multiple locations where the Cerner EHR system is being used. The merger is not complex since the vendor system is the same and the backend database structure is also similar. Nevertheless, it has its limitations when assigning a unique patient ID. The Unique ID of the patient is based on multiple identifying factors, two IDs may be generated for the same patient, and records are partially assigned to each based on location.

The CDC, ADA, Budget office, AHA, and HIMMS all come with statistics on health care settings. Their reporting standards are not same, and it is not easy to match them all into a single uniform dataset. So, we need to be careful with the reported numbers especially doing econometric studies and comparing them with prior studies.

When data from different vendors are merged, then to maintain the details, the information which is not recorded is shown as missing and this should be handled as detailed in the Missing values section. Care should be taken that the units of measurement are same when integrating data from multiple countries. Height and weight may be recorded in inches and pounds or centimeters and kilograms. It is necessary to ascertain that the measurement units are same if not than conversions should be done at the time of integration. The date may be stored or reported in different formats such as MM/DD/YYY, DD/MM/YYYY or YYYY/MM/DD, while merging a standard format should be adopted.

Data integration from disparate data sources to obtain new and informative datasets is a novel methodology and is encouraged by the research community. Srivastava, Ayyalasomayajula, Bao, Ayabakan, and Delen (2022) in their research on EHR strategy and user satisfaction have done so. They have merged the Healthcare Information and Management Systems Society (HIMSS) database with a sentiment score calculated by web scrapping online reviews posted on Glassdoor.com.

4.3.3 DATA TRANSFORMATION

This step involves Generalization, Normalization, Attribute selection, and Aggregation.

In many of the diagnostic tests performed, the results are indicated in numerical values, brief text, and or categorical values such as Low, High or Average. The machine learning algorithms can handle numerical values, and hence when using the categorical variables, we need to convert them into numerical representations. One Hot coding is a methodology to convert categorical variables for machine learning algorithms. Python and other languages provide inbuild library function for One Hot coding.

Payer information which is a non-medical variable in EHR, and should be very simple to record and use. However, in our data, we had 20 different values for the same. After multiple iterations, we reduced it to Self, Insurance, and Government as the consideration of availing the services and bargaining differs among these three groups. For this variable, we did not look at the frequency or size of these organizations but how they impact the treatment methodology. Probably the grouping would be different if we were using the same data for further analysis. We could group the same as within the network or outside as we know the rates charged are different, and it would be a good grouping in studies where medical services accessibility is the primary concern.

Normalizing is the method to scale the numerical data within the range -1.0 to 1.0 this done as the machine learning algorithms are more effective if all variable is within a specific range. Normalization is done by subtracting the minimum and then dividing by the range of the data. Python and other machine learning languages provide inbuilt standard function to normalize the attributes. In econometric and other studies, the Log Transformation is used as the range is very large, but in healthcare it is not necessary and should be avoided.

4.3.4 DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA), Discretization, Attribute subset selection are some of the methods employed to achieve reduction in dimensionality. PCA is widely used in exploratory data analysis and for predictive modeling. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The math behind its calculation is very complex, and most statistical programs, including python, provide a simple function to perform PCA.

Numerical data may have a highly skewed or non-standard distribution, which may deteriorate the performance of many machine learning algorithms. The discretization transform provides a simple method to modify a numeric input variable to have a different data distribution. The modified variable can improve the performance of the machine learning predictive models. Normally age is binned based on the context of study.

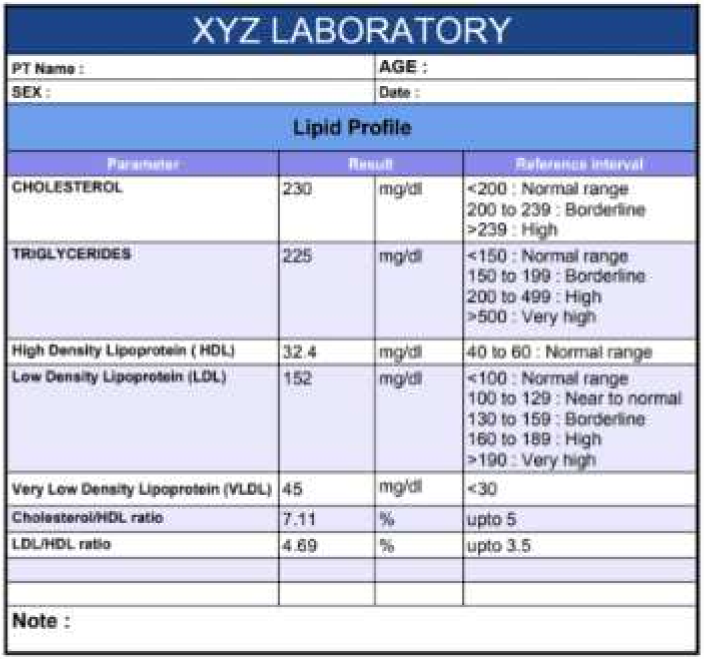
In the healthcare domain, knowledge plays a critical role in this activity. In the case of medications, we have multiple brand names for the same chemical composition. The physician may prescribe the strength, duration, and dosage of the medicine based on the treatment plan. There exists a group of medications for the same treatment, which give similar results among which one is prescribed based on multiple interaction factors.

The EHR data has several variables with high dimensionality. Medications, procedures, and all other categorical variables are very high dimensional. We have to utilize different processes and domain knowledge to reduce dimensionality. In the case of diabetes medication prescribed to our patient population, it stands at a huge number. We are not considering the prescription's strength, dosage, or duration yet. To reduce the dimensionality, we can utilize the existing clustering methods or the expert domain knowledge.

Age can be represented is recorded as number of years and can be used as such. In some studies, it may be binned into different age groups such as Kids, Adolescents, Teenagers, Adults and seniors. Binning of age may increase the dimensionality but is needed to capture the study relevance.

Multiple lab tests are performed to diagnose and study the progression of a disease. Each test reports on various parameters which may or may not be relevant or significant for the specific diagnosis. The reported results of these parameters may be numerical, categorical or brief texts. Some of the parameters are very significant others are not. Most of these reports present a reference range too. Instead of using all the reported parameters, based on domain knowledge only relevant or essential parameters should be used. We can reduce the data further by specifying normal/abnormal or Low, Normal, and High. We can further aggregate this at the test level too. This aggregation would remove the granularity but decrease the dimensionality. We can do so based on the computing power and previous study results.

Let us consider the case of lipid profile, a panel of blood tests used to find abnormalities in lipids, such as cholesterol and triglycerides. The results of this test are most often used to determine approximate risks for cardiovascular disease. They are performed frequently to assess the comorbidity in the case of a diabetic patient. A sample report is presented in figure 4.3. We can use the LDL/HDL ratio or aggregate based on domain knowledge based on assigning +1 to high and -1 to low ranges. At the patient level, we can aggregate the number of times this test was conducted and how many times it was abnormal.



XYZ LABORATORY

PT Name : AGE :

SEX : Date :

Lipid Profile

Parameter	Result		Reference interval
CHOLESTEROL	230	mg/dl	<200 : Normal range 200 to 239 : Borderline >239 : High
TRIGLYCERIDES	225	mg/dl	<150 : Normal range 150 to 199 : Borderline 200 to 499 : High >500 : Very high
High Density Lipoprotein (HDL)	32.4	mg/dl	40 to 60 : Normal range
Low Density Lipoprotein (LDL)	152	mg/dl	<100 : Normal range 100 to 129 : Near to normal 130 to 159 : Borderline 160 to 189 : High >190 : Very high
Very Low Density Lipoprotein (VLDL)	45	mg/dl	<30
Cholesterol/HDL ratio	7.11	%	upto 5
LDL/HDL ratio	4.69	%	upto 3.5

Note :

Figure 4.3 Sample Lab test report for Lipid Profile

In the case of **medicines**, domain expertise can be used to group them. They can be grouped based on the primary chemical composition, whereby all the multiple brand names are treated equivalent for the same product. This may not be a suitable methodology in econometrics or business problem studies, as dollar values are quite different. Another system is grouping the medicines based on the method of action. In the case of diabetes, it can be supplementing insulin, increasing the production of insulin, increasing the absorption of insulin, delaying the action of insulin and others. Linsky and Simon (2013) have found that medication discrepancies occurred in many cases in a system with a well-established EHR linked to pharmacy dispensing. Hence, the presumption that all the medicines have been adequately administered should not be considered.

Whenever the data pertains to a long period or through multiple different sources we face this issue of standardizing the data. Because the definition and procedure to collect the data may be different at the time of collection. This is inevitable in a activities like the decennial census of the governments. Based on the technology and the current situation new questions are added, old questions are dropped and sometimes the definition is changed. Similarly, the International Classification of Diseases has gone through multiple iterations of coding. It had gone through ICD-6, ICD-7, ICD-8a, ICD-9, and ICD-10.

To address this issue of multiple versions data should be limited to a specific version or convert the coding to a base version. However, you need to match them for longitudinal studies and select the least granular version. Horsky, Drucker, and Ramelson (2018) found that ICD coding is inconsistent among establishments based on the EHR system used. They found that the Crohn's disease and diabetes scenarios had the highest inappropriate coding and code variation rate. Codes for immunization, dialysis dependence, and nicotine dependence were often omitted.

In such scenarios we should transform the data to a coarse or higher level of versioning. In our case the ICD-9 had less classification for comorbidity details in comparison to ICD-10 so converted the ICD-10 coding to ICD-9 even though the ICD-10 was much more detailed and meaningful. ICD coding is not highly specific and definitive. The same disease may be coded differently by different practitioners at the lowest levels. Medical scribes are not physicians and hence may code the same condition differently due to paucity of knowledge or information. Billing considerations may encourage the ICD coding team to code differently to achieve maximum approval rate and insurance claim amount.

4.3.5 LONGITUDINAL VARIABLES CREATION

Marital Status of persons may change that i**s** married, divorced, separated, or widowed during the course of treatment. The probability of status is high especially when the treatment duration is long. In such scenarios, which status is to be selected. Most of the time, it would be an excellent point to consider the starting status if it is known as a causal agent, as the case is in depression and other behavioral disorders. If it is known to be an important factor in treatment such as diabetes, then being with a partner should be considered, and more accurately, the cohabiting period should be recorded.

Age should not be a problem as the date of birth is accurately recorded in the EHR. If the treatment period is more than a year, which age should be considered, diagnosed age or the current age when the analysis is done. This decision should be based on the problem being addressed. In case of progression of disease or comorbidities, the current age of the person and the age of the disease is essential. In case of severity or mortality is considered, then the age when the diagnosis was made should be considered. It is not rare to find ages recorded as 159 due to a mistake in the century column, such entries should be rectified by subtracting 100.

Insurance status is a significant factor in the case of health outcomes in the USA. In countries such as Canada, the United Kingdom, Netherland, and Sweden, where healthcare is a social welfare, and a quality service is provided free to every citizen it does not matter. In the USA, it matters a lot, and the status might change within the treatment period if it is long or the patient is visiting multiple hospitals. Hence the status which was prevalent most of the time should be considered and in case of future predictions are made, then the last status should be considered.

4.3.6. TEXTUAL DATA

Text Analytics could be used to detect the types of physical tests performed by the physician and the results reported thereof. These observations can be very useful where standard lab procedure (CPT codes) are not available such as the foot examination in the case of diabetic patients.

Clinical notes are very cryptic as physicians have less time to type or dictate while interacting with the patient. In most cases, the physician dictates while interacting with the patient, and later a medical scribe transcribes the audio recording. This may lead to errors as the medical scribes are not as knowledgeable as the physician themselves. Since the scribe is paid for the transcriptions by the duration of recording, they may use abbreviations to save time and increase productivity. The dictated clinical notes are conversational, so the grammar and sentence formation may not be complete. These notes are for the doctor's self-reference, they were not recorded for third party perusal and research. It may have extensive medical jargon and abbreviations. First, the abbreviations should be expanded based on medical dictionary, and spelling mistakes should be resolved.

The patients' social media posts in a non-clinical environment are a wealthy source for Healthcare text analytics. The first essay utilized patients' discussions on an online UK diabetes support website. It may have spelling mistakes, multiple words, or expressions to describe the symptom. Different general terminology may be used for a precise medical term. Most of the time, the patients may not be able to define the situation to medical professionals accurately. In the case of medicines, they may use the generic name, brand name, or some other slang name of it. All these considerations should be considered while preprocessing the social media text.

4.4 RESULTS AND DISCUSSION

Based on the perusal of multiple healthcare analytics research articles and the experience in analyzing the previous two essays, data preprocessing is essential and different in the case of Healthcare data. During data cleaning, the selection of values for the socio-demographic variables Race, Residency, Obesity, Smoking, and Alcoholism should be considered based on the limitations and assumptions of the data recording and collection method. Imputation of missing values should be avoided, and when done, a dummy indicator of missingness should be included in the analysis. Reasons for missingness should be identified, and accordingly, the imputation method is selected.

Data integration from multiple sources and different metrics would increase the reliability of the results while adding complexity to the preprocessing. The units of measure, date formats, and other codes should be translated to a uniform standard. While merging data from disparate sources, the timeline and measurements should be considered.

Data transformation is substantial activity in health data preprocessing. Normalizing or standardization of numerical values should be based on prior known distributions and domain knowledge. The EHR data is very high dimensional, and its reduction is necessary to run machine learning models. Aggregation, discretization, and profiling are methods that are suitable for this sector. The EHR data can consist of an extensive period of records. Hence, socio-demographic variables change over time. The proper selection of values for Marital Status, Age, and Insurance status should be based on the questions being answered.

Textual data processing in healthcare presents additional challenges in healthcare. Technical jargon, ICD codes, and abbreviations should be taken care of based on domain knowledge. The clinical notes may be transcribed inaccurately and not recorded for third-party use. Hence the validity of the text data should also be verified.

The domain knowledge of the problem area is crucial for data preprocessing in healthcare. Data with erroneous values should not be dropped. The values of other variables should be rectified or aggregated at a higher level to ignore the ambiguity. While integrating data, the data schema, units of measure, and the coding methodology should be considered. Dimensionality reduction should not be made using standard methods. It should be done by aggregation and variable selection mainly.

V. X. Liu, Bates, Wiens, and Shah (2019) developed a model and studied in the simplest terms, a number needed to benefit contextualizes the numbers needed to screen and treat, offering an opportunity to estimate the value of a clinical predictive model in action. This model goes beyond the preprocessing and does a model development and specifies the number of observations needed at each stage to be effective overall.

Business understanding of the domain knowledge is crucial for Analytics projects, as highlighted by the CRISP-DM flow diagram. This is much more necessary when non-health professionals undertake Healthcare analytics. The cost of failure and the advantage of even a marginal improvement is not known to everyone. The data collection and recording are not standard as precise as in other business processes.

To understand the available data, descriptive stats along with visualizations are relevant. Comparing these stats to prior studies and general population records is necessary for the authenticity of the research. CDC publishes annual reports for Healthcare in general and specific reports for many disease streams. Similarly, governments in each country provide annual reports. The sample data should be verified against those published reports to represent the population adequately.

Data preparation should be done based on the best practices of the industry and following the prior research. It should incorporate the knowledge gained in the previous two steps. This is the most crucial step, and innovative methods should be formed or adopted. Ultimately the accuracy and relevance of the Healthcare Analytics results arise from good Data Preparation.

CHAPTER V

SUMMARY AND CONCLUSIONS

With the adoption of EHR systems by most of the healthcare organizations and advances in data science in the past few decade Healthcare analytics has emerged as a new sub-field utilizing data analytics approaches intending to improve health care processes and better outcomes for the patients. Online social support groups for many diseases have made available the discussion topics and contents for the healthcare professional and researchers to understand the apprehensions and needs of the patients and their domestic care givers and close family members.

One of the areas of health care that has been constantly targeted for improvement worldwide for many decades is the early detection and prediction of diseases and comorbidity in the case of chronic diseases. Traditionally, diseases are detected when a patient feels the symptoms and visits a primary care, the physician identifies and confirms the disease after multiple diagnostic tests. In many cases, by this time, the disease is in an advance stage and needs advanced treatment compared to being detected at an earlier stage, this is especially true in case of diabetes. Data analytics and machine learning methods have significant potential to improve the management of diabetes by predicting comorbidities and identifying the gaps in the diabetes education programs and books. These efforts would lead to more timely *detection*, more accurate *prediction of comorbidities* and more effective *education* of Diabetes patients.

5.1 CONTRIBUTIONS

In the first essay, the Text analytics approach, along with web scrapping, has been extended to diabetes management to identify patients' unmet informational needs and to improve patient education programs. While the books, websites, and Diabetes Education Programs individually impart expert education on the topics they are addressing, they may not address all the informational needs of the patients. So, the question is how to broadly identify some of the unmet questions in each topic and assist the experts in designing and developing the educational programs to fill the gaps. By analyzing the thirty thousand threads, we identified the five broad discussion topics. In each topic, we identified the significant questions being asked. We explained how these questions are answered and explained in the available books and educative material on professional and governmental websites. The results explain the patients' inquiries and how they might be addressed to avoid the gaps.

The second essay provides a framework to use standard EHR data to rank-order potential comorbidities on the likelihood of developing comorbidity. While there are studies to identify the risk factors for each of the comorbidities, rank ordering and identifying the risk factors for a diabetes patient towards all the comorbidities was not empirically studied. The rank-ordering would assist the physicians in prioritizing the top comorbidities and avoid missing regular screening where necessary. The ANN model was applied to EHR data containing sociodemographic, medication, and laboratory tests to rank-order the comorbidities. The results show a high probability of accurately identifying the difficult to predict and rare comorbidities.

In the third essay, we have perused the healthcare analytics literature and introspected the data preprocessing steps necessary in the previous two studies. The idea of this study was to use the known data preprocessing studies and identify how they were modified in the case of healthcare analytics. Our exploration showed that the standard preprocessing could not be applied to EHR and other health-related data without relevant modification. Having in-depth knowledge of the subject would assist in correctly identifying the methodology to be applied.

5.2 ASSUMPTIONS AND LIMITATIONS

The present work involves several limitations, as discussed below.

In the first essay, we limited our sample to social media posts in popular UK diabetics forums, assuming that the concerns raised in the US and other countries would be similar. Identifying the type of diabetics' education and level of training the forum posters have gone through would have been time-consuming, difficult and infeasible. Finding a similar forum in the USA with such an active participating member was not easy. In the USA, due to HIPPA regulations, it would be difficult for the forum to store and display patient's information. We also implicitly assumed that the patients had the diabetes management education training and were inquiring about topics after doing due diligence. It is also assumed that diabetic management education is similar in content worldwide, or due to the prevalence of the internet, people could access quality education online from multiple sources and countries.

In the second study, we assumed that the EHR data provided by the Cerner Health facts represent the actual population of the US, as it was substantial and comprehensive. It is impossible to verify whether the patients consumed all the medications per the dosage prescription and vice versa. Due to billing and other insurance limitations, some tests performed may not be recorded. The ICD codes assigned may have been not accurate by error or reasons of insurance claim approval procedure. The patients might have treatment for some duration outside the Cerner network, and we may have partial records of a person's medical history.

Moreover, a limitation to the third study is that we are considering the detailed preprocessing steps conducted in these two studies. The prior published healthcare research might have reported only a partial preprocessing step. It is common in research articles to delve more into the theoretical and methodological issues in comparison to technical and procedural steps. A limited set of articles were perused due to time constraints. Most of the studies used MIMIC and other public data sets that are well-curated and structured for research.

5.3 FUTURE RESEARCH DIRECTIONS

This work leads to several research areas in diabetes education and management, as discussed next.

a) Compare and contrast diabetes education in different countries based on the social media forums inquires. Patients can access the books, websites, and other diabetes educational material from other countries. The healthcare delivery system is different and cannot be widely accessed due to national boundaries and cost limitations. (National, Private Insurance, Mixed) effects the education.

b) Test the comorbidity results using the EHR data from EPIC or other EHR systems to validate the generalizability over the population of the US. Improve the accuracy by incorporating lifestyle, location, income and education, and other variables.

c) Develop a web application to predict the comorbidities in the case of diabetic patients for medical professionals and patients. The web app may be similar to credit scoring with minimal data. It would rank the comorbidities and keep track of the prognosis.

d) Systematic literature review using PRISMA guidelines on data preprocessing for Healthcare. As Payrovnaziri et al. (2020) observed in their review, XAI evaluation in medicine has not been adequately and formally practiced. Reproducibility remains a critical concern as the data is mainly protected and not public.

References

Abidi, S. S. R., Singh, A. K., & Christie, S. (2017) Transcription of case report forms from unstructured referral letters: A semantic text analytics approach. In*: Vol. 228* (pp. 322-326): IOS Press.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87-97. doi:10.1016/j.dss.2012.04.005

Alder, S. (2022, Jan 28, 2022). What is Considered PHI Under HIPAA? Retrieved from https://www.hipaajournal.com/CONSIDERED-PHI-HIPAA/

Altshuler, A., & Ladd, D. L. (2013). *The Medical Library Association Guide to Finding Out About Diabetes: The Best Print and Electronic Resources (Guide to Finding Out About Diabetes)*: American Library Association.

American Diabetes, A. (2019). 4. Comprehensive Medical Evaluation and Assessment of Comorbidities: &lt;em&gt;Standards of Medical Care in Diabetes—2019&lt;/em&gt;. *Diabetes Care, 42*(Supplement 1), S34. doi:10.2337/dc19-S004

AMERICAN DIABETES ASSOCIATION. (2004). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care, 27*(suppl 1), s5. doi:10.2337/diacare.27.2007.S5

American Diabetes Association. (2018). Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care, 41*, 917-928.

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-Based Attribution Methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 169-191). Cham: Springer International Publishing.

Association, A. D. (1999). *Complete guide to diabetes : the ultimate home diabetes reference* (2nd ed. ed.). Alexandria, Va: American Diabetes Association.

Association, A. D. (2010). *Diabetes A to Z : what you need to know about diabetes, simply put* (Sixth edition. ed.). Alexandria, Va: American Diabetes Association.

Ayyar, S., & IV, O. B. D. t. W. (2016). *Tagging patient notes with icd-9 codes*. Paper presented at the Proceedings of the 29th Conference on Neural Information Processing Systems.

Bajor, J. M., & Lasko, T. A. (2017). *PREDICTING MEDICATIONS FROM DIAGNOSTIC CODES WITH RECURRENT NEURAL NETWORKS*. Paper presented at the ICLR 2017.

Balakrishnan, V., Shakouri, M. R., Hoodeh, H. J. J. o. I., & Systems, F. (2013). Developing a hybrid predictive system for retinopathy. *25*(1), 191-199.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., . . . Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. doi:https://doi.org/10.1016/j.inffus.2019.12.012

Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. e. (2017). *Multi-Label Classification of Patient Notes a Case Study on ICD Code Assignment*. Paper presented at the Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Brahma, A., Goldberg, D. M., Zaman, N., & Aloiso, M. (2021). Automated mortgage origination delay detection from textual conversations. *Decision Support Systems, 140*. doi:10.1016/j.dss.2020.113433

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission*. Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia. https://doi.org/10.1145/2783258.2788613

Cassim, N., Mapundu, M., Olago, V., George, J. A., & Glencross, D. K. (2019) Using big data techniques to improve prostate cancer reporting in the Gauteng province, South Africa. In*: Vol. 264* (pp. 1437-1438): IOS Press.

Caucheteur, D., Gobeill, J., Mottaz, A., Pasche, E., Michel, P. A., Mottin, L., . . . Ruch, P. (2020) Text-mining services of the Swiss variant interpretation platform for oncology. In*: Vol. 270* (pp. 884-888): IOS Press.

Celis, L. E., Keswani, V., & Vishnoi, N. (2020). *Data preprocessing to mitigate bias: A maximum entropy based approach*. Paper presented at the Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v119/celis20a.html

Centers for Disease Control Prevention, A., GA: US Department of Health, & Services, H. (2020). National diabetes statistics report, 2020. 12-15.

Centers for Medicare & Medicaid Services. (2021). Retrieved from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical

Chatterjee, S., Deng, S., Liu, J., Shan, R., & Jiao, W. (2018). Classifying facts and opinions in Twitter messages: a deep learning-based approach. *Journal of Business Analytics, 1*(1), 29-39. doi:10.1080/2573234X.2018.1506687

Chen, R., Zheng, Y., Xu, W., Liu, M., & Wang, J. (2018). Secondhand seller reputation in online markets: A text analytics framework. *Decision Support Systems, 108*, 96-106. doi:10.1016/j.dss.2018.02.008

Chitravathi, R., & Kanimozhi, G. (2019). Disease prediction using Snn over big data. *International Journal of Innovative Technology and Exploring Engineering, 8*(10), 1744-1749. doi:10.35940/ijitee.J9107.0881019

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). *Doctor AI: predicting clinical events via recurrent neural networks*. Paper presented at the Proceedings of Machine Learning for Healthcare 2016.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association, 24*(2), 361-370. doi:10.1093/jamia/ocw112

Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2019). Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association, 27*(1), 39-46. doi:10.1093/jamia/ocz101 %J Journal of the American Medical Informatics Association

Clinic, M. (2009). *Mayo Clinic : the essential diabetes book : [how to prevent, control and live well with diabetes]* (1st edition. ed.). New York: Time Home Entertainment.

Delen, D. (2021). *Predictive analytics: Data mining, machine learning and data science for practitioners, 2nd edition* (2 ed.). Upper Saddle River, NJ: Pearson FT Press.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. J. J. o. t. A. M. I. A. (2017). De-identification of patient notes with recurrent neural networks. *Journal ofthe American Medical Informatics Association, 24*(3), 596-606. doi:10.1093/jamia/ocw156

Desai, D., Mehta, D., Mathias, P., Menon, G., & Schubart, U. K. (2018). Health Care Utilization and Burden of Diabetic Ketoacidosis in the U.S. Over the Past Decade: A Nationwide Analysis. *Diabetes Care, 41*(8), 1631. doi:10.2337/dc17-1379

Dutta, A., Batabyal, T., Basu, M., & Acton, S. T. (2020). An efficient convolutional neural network for coronary heart disease prediction. *Expert Systems with Applications, 159*, 113408. doi:https://doi.org/10.1016/j.eswa.2020.113408

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research, 9*.

Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems, 112*, 88-97. doi:10.1016/j.dss.2018.06.008

Geiss, L. S., Li, Y., Hora, I., Albright, A., Rolka, D., & Gregg, E. W. (2019). Resurgence of Diabetes-Related Nontraumatic Lower-Extremity Amputation in the Young and Middle-Aged Adult U.S. Population. *Diabetes Care, 42*(1), 50. doi:10.2337/dc18-1380

Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2016). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association, 24*(1), 198-208. doi:10.1093/jamia/ocw042 %J Journal of the American Medical Informatics Association

Gruss, R., Abrahams, A. S., Fan, W., & Wang, G. A. (2018). By the numbers: The magic of numerical intelligence in text analytic systems. *Decision Support Systems, 113*, 86-98. doi:10.1016/j.dss.2018.07.004

Hassan, N. R. (2019). The origins of business analytics and implications for the information systems field. *Journal of Business Analytics, 2*(2), 118-133. doi:10.1080/2573234X.2019.1693912

Hassan Zadeh, A., & Jeyaraj, A. (2018). Alignment of business and social media strategies: insights from a text mining analysis. *Journal of Business Analytics, 1*(2), 117-134. doi:10.1080/2573234X.2019.1602002

Hawley, J. A., & Gibala, M. J. (2012). What's new since Hippocrates? Preventing type 2 diabetes by physical exercise and diet. *Diabetologia, 55*(3), 535-539. doi:10.1007/s00125-012-2460-1

Horsky, J., Drucker, E. A., & Ramelson, H. Z. (2018). Accuracy and Completeness of Clinical Coding Using ICD-10 for Ambulatory Visits. *AMIA ... Annual Symposium proceedings. AMIA Symposium, 2017*, 912-920.

Ibrahim, N. F., & Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems, 121*, 37-50. doi:10.1016/j.dss.2019.03.002

Idri, A., Benhar, H., Fernández-Alemán, J. L., & Kadi, I. (2018). A systematic map of medical data preprocessing in knowledge discovery. *Computer Methods and Programs in Biomedicine, 162*, 69-85. doi:https://doi.org/10.1016/j.cmpb.2018.05.007

Jain, A. G., Guan, J., FaisalUddin, M., Manoucheri, M., & Fang, C. (2019). Improving breast cancer screening rates in a primary care setting. *The Breast Journal, 25*(5), 963-966. doi:10.1111/tbj.13377

Kam, H. J., & Kim, H. Y. (2017). Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine, 89*, 248-255. doi:10.1016/j.compbiomed.2017.08.015

Khan, S., & Shamsi, J. A. (2018). Health Quest: A generalized clinical decision support system with multi-label classification. *Journal of King Saud University - Computer and Information Sciences*. doi:https://doi.org/10.1016/j.jksuci.2018.11.003

Lee, E., & Zhao, H. (2020). Deriving topic-related and interaction features to predict top attractive reviews for a specific business entity. *Journal of Business Analytics, 3*(1), 17-31. doi:10.1080/2573234X.2020.1768808

Lehne, M., Luijten, S., Vom Felde Genannt Imbusch, P., & Thun, S. (2019) The use of FHIR in digital health – A review of the scientific literature. In*: Vol. 267* (pp. 52-58): IOS Press.

Levin, M. E., & Pfeifer, M. A. (1998). *The uncomplicated guide to diabetes complications*. Alexandria, Va: American Diabetes Association.

Levine, J. A. (2011). Poverty and Obesity in the U.S. *60*(11), 2667-2668. doi:10.2337/db11-1118 %J Diabetes

Li, Y., & Ngom, A. (2015, 9-12 Nov. 2015). *Data integration in machine learning*. Paper presented at the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Linsky, A., & Simon, S. R. (2013). Medication discrepancies in integrated electronic health records. *BMJ Quality &amp;amp; Safety, 22*(2), 103. doi:10.1136/bmjqs-2012-001301

Liu, V. X., Bates, D. W., Wiens, J., & Shah, N. H. (2019). The number needed to benefit: estimating the value of predictive analytics in healthcare. *Journal of the American Medical Informatics Association, 26*(12), 1655-1659. doi:10.1093/jamia/ocz088

Liu, Z., Tang, B., Wang, X., & Chen, Q. J. J. o. b. i. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics, 75*, S34-S42.

Ljubic, B., Hai, A. A., Stanojevic, M., Diaz, W., Polimac, D., Pavlovski, M., & Obradovic, Z. (2020). Predicting complications of diabetes mellitus using advanced machine learning algorithms. *Journal of the American Medical Informatics Association, 27*(9), 1343-1351. doi:10.1093/jamia/ocaa120

Maragatham, G., & Devi, S. (2019). LSTM Model for Prediction of Heart Failure in Big Data. *Journal of Medical Systems, 43*(5), 111. doi:10.1007/s10916-019-1243-3

Martin, B.-J., Chen, G., Graham, M., & Quan, H. J. B. h. s. r. (2014). Coding of obesity in administrative hospital discharge abstract data: accuracy and impact for future research studies. *14*(1), 1-8.

Mayo Clinic. (2021). Diabetes , Complications. Retrieved from https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

McHaney, R., Tako, A., & Robinson, S. (2018). Using LIWC to choose simulation approaches: A feasibility study. *Decision Support Systems, 111*, 1-12. doi:10.1016/j.dss.2018.04.002

McVeigh, K. H., Newton-Dame, R., Chan, P. Y., Thorpe, L. E., Schreibstein, L., Tatem, K. S., . . . Perlman, S. E. (2016). Can Electronic Health Records Be Used for Population Health Surveillance? Validating Population Health Metrics Against Established Survey Data. *EGEMS (Washington, DC), 4*(1), 1267-1267. doi:10.13063/2327-9214.1267

Miller, A. N., Bharathan, A., Duvuuri, V. N. S., Navas, V., Luceno, L., Zraick, R., . . . Thrash, K. (2022). Use of seven types of medical jargon by male and female primary care providers at a university health center. *Patient Education and Counseling, 105*(5), 1261-1267. doi:https://doi.org/10.1016/j.pec.2021.08.018

Miller, R. G., Costacou, T., & Orchard, T. J. (2019). Risk Factor Modeling for Cardiovascular Disease in Type 1 Diabetes in the Pittsburgh Epidemiology of Diabetes Complications (EDC) Study: A Comparison With the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study (DCCT/EDIC). *Diabetes, 68*, 409-419.

Miller, T. J. A. I. (2019). Explanation in artificial intelligence: Insights from the social sciences. *267*, 1-38.

Mirkes, E. M., Coats, T. J., Levesley, J., & Gorban, A. N. (2016). Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine, 75*, 203-216. doi:https://doi.org/10.1016/j.compbiomed.2016.06.004

Payne, T. H., Corley, S., Cullen, T. A., Gandhi, T. K., Harrington, L., Kuperman, G. J., . . . Zaroukian, M. H. (2015). Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. *Journal of the American Medical Informatics Association, 22*(5), 1102-1110. doi:10.1093/jamia/ocv066 %J Journal of the American Medical Informatics Association

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., . . . He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association, 27*(7), 1173-1185. doi:10.1093/jamia/ocaa053

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics, 69*, 218-229. doi:https://doi.org/10.1016/j.jbi.2017.04.001

Piette, J. D., & Kerr, E. A. (2006). The Impact of Comorbid Chronic Conditions on Diabetes Care. *Diabetes Care, 29*(3), 725. doi:10.2337/diacare.29.03.06.dc05-2078

Piri, S. (2020). Missing care: A framework to address the issue of frequent missing values;The case of a clinical decision support system for Parkinson's disease. *Decision Support Systems, 136*, 113339. doi:https://doi.org/10.1016/j.dss.2020.113339

Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems, 106*, 15-29. doi:10.1016/j.dss.2017.11.006

Piri, S., Delen, D., Liu, T., & Zolbanin, H. M. (2017). A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. *Decision Support Systems, 101*, 12-27. doi:https://doi.org/10.1016/j.dss.2017.05.012

Pitt, M. B., & Hendrickson, M. A. J. J. o. g. i. m. (2021). Response to Letter to the Editor Re: Eradicating Jargon-Oblivion—a Proposed Classification System of Medical Jargon. *36*(4), 1112-1112.

Rasmy, L., Wu, Y., Wang, N., Geng, X., Zheng, W. J., Wang, F., . . . Zhi, D. (2018). A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *Journal of Biomedical Informatics, 84*, 11-16.

Reading Turchioe, M., Grossman, L. V., Myers, A. C., Baik, D., Goyal, P., & Masterson Creber, R. M. (2020). Visual analogies, not graphs, increase patients' comprehension of changes in their health status. *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocz217

Saba, T. J. M. R., & Technique. (2021). Computer vision for microscopic skin cancer diagnosis using handcrafted and non‐handcrafted features. *84*(6), 1272-1283.

Samuelson, D. A., Spirer, H. F. J. H. r., & straight, s. G. t. r. (1992). Use of incomplete and distorted data in inference about human rights violations. 62-77.

Seker, E., Talburt, J. R., Greer, M. L. J. S. i. H. T., & Informatics. (2022). Preprocessing to Address Bias in Healthcare Data. *294*, 327-331.

Shaw, R. J., Yang, Q., Barnes, A., Hatch, D., Crowley, M. J., Vorderstrasse, A., . . . Steinberg, D. (2020). Self-monitoring diabetes with multiple mobile health devices. *Journal of the American Medical Informatics Association, 27*(5), 667-676. doi:10.1093/jamia/ocaa007

Sheehan, J., & Ulchaker, M. M. (2012). *Obesity and type 2 diabetes mellitus*. Oxford ;: Oxford University Press.

Singh, N., & Varshney, U. (2020). IT-based reminders for medication adherence: systematic review, taxonomy, framework and research directions. *European Journal of Information Systems, 29*(1), 84-108. doi:10.1080/0960085X.2019.1701956

Skevofilakas, M., Zarkogianni, K., Karamanos, B. G., & Nikita, K. S. (2010). *A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus*. Paper presented at the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology.

Song, X., Waitman, L. R., Hu, Y., Yu, A. S. L., Robins, D., & Liu, M. (2019). Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *Journal of the American Medical Informatics Association, 26*(3), 242-253. doi:10.1093/jamia/ocy165 %J Journal of the American Medical Informatics Association

Srivastava, A., Ayyalasomayajula, S., Bao, C., Ayabakan, S., & Delen, D. (2022). Relationship between electronic health records strategy and user satisfaction: a longitudinal study using clinicians' online reviews. *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocac082

Summers-Gabr, N. M. J. P. T. T., Research, Practice,, & Policy. (2020). Rural–urban mental health disparities in the United States during COVID-19. *12*(S1), S222.

Tao, J., Deokar, A. V., & Deshmukh, A. (2018). Analysing forward-looking statements in initial public offering prospectuses: a text analytics approach. *Journal of Business Analytics, 1*(1), 54-70. doi:10.1080/2573234X.2018.1507604

Tuch, B., Dunlop, M., & Proietto, J. (2000). *Diabetes research : a guide for postgraduates*. Australia: Harwood Academic.

Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association*. doi:10.1093/jamia/ocz130

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87-95. doi:10.1016/j.dss.2017.11.001

Wukich, D. K., Raspovic, K. M., & Suder, N. C. (2017). Patients With Diabetic Foot Disease Fear Major Lower-Extremity Amputation More Than Death. *Foot & Ankle Specialist, 11*(1), 17-21. doi:10.1177/1938640017694722

Yuan, H., Lau, R. Y. K., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems, 91*, 67-76. doi:10.1016/j.dss.2016.08.001

Zhang, Y., Chen, R., Tang, J., Stewart, W. F., & Sun, J. (2017). *Leap: Learning to prescribe effective and safe treatment combinations for multimorbidity*. Paper presented at the proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining.

Zoega, G. M., Gunnarsdóttir, Þ., Björnsdóttir, S., Hreiðarsson, Á. B., Viggósson, G., & Stefánsson, E. (2005). Screening compliance and visual outcome in diabetes. *Acta Ophthalmologica Scandinavica, 83*(6), 687-690. doi:10.1111/j.1600-0420.2005.00541.x

VITA

SURYA BHASKAR AYYALASOMAYAJULA

Candidate for the Degree of

Doctor of Philosophy

Dissertation: USING MACHINE LEARNING METHODS TO IMPROVE HEALTHCARE DELIVERY IN DIABETES MANAGEMENT

Major Field: Business Administration (MSIS)

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Business Administration (MSIS) at Oklahoma State University, Stillwater, Oklahoma in July, 2022.

Completed the requirements for the Master of Science in Business Analytics at Oklahoma State University, Stillwater, Oklahoma in 2017.

Completed the requirements for the Bachelor of Science in Sciences (Computer) at University of Delhi, Delhi, India in 1987.

Experience:

Graduate Teaching Associate at the MSIS Department, Oklahoma State University, Stillwater, Oklahoma, 2017-2022.

Graduate Teaching Assistant at Watson Graduate College, Oklahoma State University, Stillwater, Oklahoma, 2016.

Software development, maintenance and customization at Bharatiya Ygyaniky, Delhi, India, 203-2015.

Professional Memberships:

Association for Information Systems (AIS), 2017-2022

Institute of Operations Research and Management Sciences (INFORMS) 2017-2022.