TEMPORAL CBR FOR PERSONALIZING HYPERTENSION TREATMENT

A Dissertation Presented

By

Niloofar Jalali

to

The Department of Mechanical and Industrial Engineering

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the field of
Mechanical Engineering

Northeastern University
Boston, Massachusetts

December 2016

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my advisors Prof. Abe Zeid and Prof. Sagar Kamarthi for the continuous support of my Ph.D program, for their patience, motivation, and immense research experiences. Their guidance helped me in all stages of research and writing of this dissertation.

Besides my advisors, I would like to thank the rest of my thesis committee: Dr. Stephen Agboola and Dr. Kamal Jethwani for their insightful comments and encouragement, but also for the hard questions which helped me to examine my research from various perspectives.

Last but not the least, I would like to thank my family: my father who inspired me the most to continue my PhD study. My mother, who always believed in me, supported me and taught me to overcome the fear of my failure and to my brother who encouraged me to have faith in myself and never stop trying.

Table of Contents

List of Tables

List of Figures

ABSTRACT

Hypertension (high blood pressure) is a serious chronic disease that is a major risk factor of cardiovascular problems, heart failures, and strokes. Hypertension disease cannot be controlled effectively with the current treatment methods due to barriers such as poor adherence to prescription caused by drugs' side effects, patients' poor tolerability to medication, and ineffectiveness of drugs.

In order to better control hypertension, different research groups have tried personalized medicine for treating the disease by using the patients' genetic information. However, this approach hasn't been successful since hypertension is less governed by genetic factors. The Joint National Committee (JNC) reports identify that patient parameters such as family history, demographic information, physical examination and laboratorial test results are significant factors for diagnosis and evaluation of hypertension disease.

In this study, a new approach based on Temporal Case Based Reasoning (CBR) has been investigated to personalize hypertension treatment. This method leverages what worked well for similar patients in the past. This approach has been built on ten years of clinical data obtained from the Mass General Hospital. The dataset has multiple medical records for each patient. Each patient's medical record corresponds to one episode (visit). It is assumed that the oldest episode of each patient is the first instance of hypertension diagnosis. After clustering the similar patients based on their oldest episodes, the subsequent episodes of all patients in each cluster are examined to find an effective treatment for the patients in that cluster. By evaluating the efficacy of the different medications prescribed to the patients in a cluster, successful treatments are selected and classified to build an adaptive treatment model for that cluster of patients. When a new patient comes in with hypertension diagnosis, his/her state (demographic, physical exam and lab results) is compared with the state of the past patients as they were at their oldest episodes to retrieve the top three similar patient matches. By combining the treatments identified for these three matching patients, an effective treatment for the new patient is synthesized. Moreover, this method could provide physicians with other useful information such as the percentage of successfully treated patients in each cluster and a shortest treatment path for desirable outcomes. The validation of results shows that, in the majority of cases, temporal CBR method recommends the right treatment. The accuracy of the prescription can be improved if the medical records have more data about the patient lifestyle parameters such as diet and exercise.

In future work, the recommendations of the proposed approach can be validated by seeking inputs from expert physicians. This approach is broadly applicable to treating other chronical diseases as well.

Chapter I: Introduction

1.1 Introduction

Hypertension is one of the most important chronic diseases that has been found recently as a major risk factor of cardiovascular problem which causes morbidity and mortality. It causes 6% of deaths worldwide, which is acknowledged as a major public health concern in a society. Due to asymptomatic nature of this disease, many hypertensive patients don't see a doctor. Therefore, diagnosis, control and treatment of hypertension is one of the major challenges in the United States.

Besides the diagnosis of this disease, the medical treatment hasn't had yielded successful outcomes. The current treatment procedures for this disease are provided according to some guidelines that make the decisions convenient for the physicians. However, this approach hasn't looked into individualized patient information to personalize the treatment. Hypertension and its associated complications increase with the age, population, density, and obesity. On the other hand, any single method for diagnosing, treating or preventing this disease wouldn't be successful for all individuals. These limitations underline the necessity of considering patient specific factors that play an important role in determining the better control and treating of hypertension. Therefore, in this study, a new approach for personalizing treatment of hypertension is proposed.

1.2 Motivation

The investigational studies of personalized medicine have been growing significantly in the healthcare area. This allows patients to benefit from better treatment and avoidance from the drugs' side effects. Generally, the idea of using personalized medicine is to use genomic information of individuals to get better idea about the patients' characteristics, such as interpreting the susceptibility level of response to different types of drugs, recognizing the significance of genetic factors to raise disease possibilities. However, there are some barriers to this approach, such as cost of investigations and also the pain associated with this procedure (Byrd 2016).

Despite of those barriers, there are some studies that attempt to personalize the treatment of hypertension based on the genetic parameters. However, no explicit results are available (Byrd 2016). Therefore, it could be concluded that, the development of hypertension is less governed by genetic factors.

There are so many factors that could cause hypertension. These factors include, people life style, diet, family history, or some other comorbidities. Based on the results in different studies, people who are having a healthy diet and exercising regularly are more unlikely to suffer from hypertension. Moreover, patients who are suffering from comorbidities such as diabetics, kidney problem and heart disease may not be going through the same treatment plans as other hypertensive patients. Therefore, in this study, the other contributing factors are considered as the predictors to personalize the hypertension medication.

1.3 Research Hypothesis

Personalized medicine for hypertension means to consider each individual's characteristics as a factor to predict the most effective drug that could lower the patient blood pressure. This approach could also reduce the side effects of some other drugs by considering the existence of other comorbidities. Hence, in this study, it is hypothesized that using patients' characteristics such as demographic information and laboratorial results would better determine the hypertension treatment than current hierarchical guidelines. This idea has been implemented through a predictive analytics technique called Case Based Reasoning (CBR).

1.4 Case Based Reasoning

The predictive data mining algorithms are known to be the most powerful tools for analyzing data and classifying. The objective of this study is to apply the case-based reasoning technique, as a predictive algorithm for treatment of hypertension. The core idea of this technique is to solve new problems by adapting existing solutions that were used for solving similar previous problems. The patient's parameters such as demographic information, life style, comorbidity diseases and laboratory test results are used as the index to measure the similarity. Then, after finding the similar cases, the related treatments are being used directly or being modified as a new treatment. The proposed approach is a new technique to personalize the treatment based on patient characteristics.

A case usually refers to a problem that was captured, learned and used to solve future similar problems. In order to match the new problem with previous cases, the cases should be indexed and interpreted correctly. Indexing can store the previous cases properly so they can be recalled easily. However, interpreting a case is the process of comparing the new problem with previous cases. The process of indexing and interpreting the cases is called case retrieval. Since the retrieved cases may not be exact the same as the new problem, they need to be adapted to meet the new problem.

Recent studies that apply CBR include time as a parameter to retrieve most effective solutions that vary with time. This approach is more helpful in the healthcare area in which one need to look at historical evidence to find an accurate diagnostic or treatment regime. Hence, in this study, a time-based CBR is applied to track the outcomes of a drug therapy on hypertensive patients and find the most effective drug as a suggested treatment. Initially, observations or episodes in each patient's medical records are chronologically ordered such that the oldest observation is placed first in episode sequence and the latest observation is placed the last. It is assumed that the first episode of each patient is the first state of diagnose; so when a new patient comes, his/her status should be compared with the first status of past patients. Therefore, the retrieval process calculates the similarity between the new patients and the most similar past patients based on the first episode of the past patients' records. Due to difference of therapies for matching patients, the best treatment couldn't be recognized without knowing the result of the treatments. Therefore, the subsequent records of matching patients should be compared to find the best treatment. This might even require using the combination of treatments from all matching patients to get the desirable result.

1.5 Significance of Personalized Treatment

Due to ineffectiveness of genetic information to improve the outcomes of hypertension treatment, other factors should be considered as significant predictors to personalize the hypertension treatment. Based on current Joint National Committee (JNC) reports, the important role of other factors to personalization of hypertension treatment has been demonstrated. However, those factors haven't been included to update the guideline and are merely acknowledging the fact that whether the patient is suffering from other diseases like diabetics or a kidney problem (Byrd 2016). Those factors are included as urine test result, blood glucose, electrocardiography, creatinine and sodium. Moreover, the stochastic results have referred to the significant function of other features' group like diet, smoking, exercise, alcohol consumption, BMI and other demographic information of patients to better address the treatment.

Therefore, due to the important role of other factors and the lack of their usage in defining the treatment procedure, the personalized treatment of hypertension is realized by using those parameters. The clinical records of hypertensive patients who are suffering from diabetics are selected. Then after cleaning the data, the laboratorial results, vital signs and demographic data have been applied by predictive algorithms.

1.6 Assumptions

In temporal CBR, the characteristics of new patient is compared to previous records in order to find the best match. It is assumed that the new patient is one who has not had any previous records and just started the treatment. It is also hypothesized that the new patient is starting his/her treatment while he/she is suffering from the level 2 or 3 of hypertension. Based on JNC report, three different levels of blood pressure can define the state of hypertension as shown in Table1. Level 1 is considered a healthy patient in this study.

Standard Levels of hypertension	**Levels**
SBP<140 & DBP<90	1
140≤SBP<160 & 90≤DBP<99	2
SBP≥160 & DBP≥100	3

Table 1. 1:Standard Levels of Hypertension

Moreover, all other previous patients are assumed to have started their first treatment while they are diagnosed of having level 2 or 3 of hypertension. After cleaning the data, the earliest record of each patient that shows the systolic blood pressure of 140 or more is selected as the first visit of the patient.

The other assumption is to neglect the time interval between each visit. For the ideal treatment, the time duration between each visit should be between two to six months. However, in reality, the clinical records of some patients are more than six months.

1.7 Limitation

In order to increase the accuracy of the personalized treatment, more information regarding the patients' status would help to identify the cause of the disease and improve the quality of the treatment. One of the important and common approaches for control and treatment of the hypertension is to change the life style, like having healthy diet, exercising regularly, stop smoking, moderating alcohol consumption, control sodium intake. Those modifications are always known as a significant step towards lowering the blood pressure which is prescribed as a treatment for early stages of hypertension with/without drug therapy. Despite the importance of these parameters, they are usually difficult to be measured. Most of them cannot be supervised by clinicians and may not be recorded suitably because of the negligence or forgetfulness of patients to provide the accurate information. Therefore, due to lack of this information, they weren't included in this study. On the other hand, given that all the patients are suffering from comorbidities, it is assumed that the life style modification is already being controlled and the study is merely considering the drug therapy as a defining factor to change the level of the blood pressure.

1.8 Summary

Based on recent reports, the rate of mortality associated with hypertension is continuously growing from 2000-2013 ("CDC: Hypertension-Related Mortality Has Climbed Since 2000" 2016). Despite of several research efforts in this area, understanding the cause of this disease and also the effectiveness of certain treatment hasn't been achieved yet. In order to better control the treatment of this disease, the idea of using personalized medicine has been considered. So, several studies have been exploring the effect of genetics on developing the hypertension and also responding to the antihypertensive drugs. However, the genome information doesn't expose the salient information regarding the blood pressure changes. It is also concluded that the DNA sequence shall not be considered as a personalization factor for hypertension treatment in future studies (Byrd 2016).

On the other hands, the current approach for hypertension treatment rely on the systematic statements as guideline to help clinicians to find the right medication. However, those recommendations may not always be accurate. Because, for testing the medication, the small group of patients cannot be used as an accurate indicator of all people. Moreover, those results are sometimes adopted and changed by clinicians in order to reach better results. However, it may lead to opposite consequences.

Therefore, in this study, the new approach of applying personalized treatment for hypertension by using patients' characteristics is introduced. In this research, a clinical dataset from Partners Health Connected group has been used. The data records include the information of hypertensive patients who were also suffering from diabetes. They contain the history visits of patients during five to ten years of their treatment. For each visit, the important parameters were measured and the prescribed medications recorded.

To facilitate the process, the important parameters of patients that have significant effect on changing the blood pressure such as demographic information, vital signs and some lab results has been selected. Then after the data were cleaned, the final records would indicate the different records of patients during their treatment process. It is hypothesized that each patient has high level of blood pressure for their first visit. That has been chosen as a reference for the new patient. Therefore, the dataset would be divided into two groups of first visit and subsequent visits of records. The subsequent visits would demonstrate the visit history of each patient after their first episode.

So, to provide the personalized treatment for the new patient, his/her characteristics is compared to the first visit's status of all other previous patients. Case-based reasoning (CBR) is used to retrieve the similar past records and adapting the successful results to discover the new treatment. One of the advantages of this approach over other methods is that the specific information of previous cases could be extracted instead of general information of problem domain. It is also known as continuous learning approach, since the new problem is saved after being solved and will be used for future similar cases (E.

Plaza 1994). The further details of this algorithm and its different applications in personalized treatment are discussed further in later chapters of this dissertation.

Chapter II: Literature Review

2.1 The State of Personalized Medicine in Diagnosis and Treatment

Based on recent studies, scientists have found that, human's genes have important role in causing some diseases. Therefore, by observing the change of genetic response to drug treatment, the treatment goal was changed to match the molecular basis of the disease. Those genes are usually identified through some tests at the early stage of the diagnosis. Therefore, the treatment of the disease could be predicted. In order to achieve desirable results for individualized treatment, the diagnostic tests should be precise and the triggering genes are identified correctly. This approach has been used successfully to ascertain the breast tumors' growth factor receptor type 2. This method was also applied in treatment of lung cancer (Hensing et al. 2014).

Despite the important effect of using genome information for better diagnosing and personalizing the treatment, they are not always a good solution since Genetic tests are not perfect. The mutations' outcome won't be always accurate. So the specificity and sensitivity consequences of diagnosis and treatment using this approach should be assessed in advance.

Despite the confirmation of National Institute Health (NIH) and former commissioner of the US Food and Drug Administration (FDA) about prominent effect of using personalized medicine, many laboratories have started complex genetic test that is hard to interpret. Therefore, there were some concerns regarding the absence of FDA supervision. On the other hand, based on NIH results, no inclusive reports regarding all genetic tests around different laboratories exist (Hamburg and Collins 2010).

2.2 Personalized Medicine for Hypertension Treatment

The idea of using personalized medicine to treat the hypertension starts with the question of whether this is a beneficial approach.

Hypertension is the major chronic disease that is most commonly observed by clinicians. The stochastic results have indicated that almost 78 million people in US are suffering from hypertension. Moreover, the National Health and Nutrition Evaluation Survey (NHANES) in 2007-2010 revealed that 85.1% of hypertensive patients are aware of their disease. While only 79.4% are being treated and 52.5% are having controlled blood pressure (Go et al. 2013). Hypertensive patients usually have other risk factors like lipid problem, diabetes, kidney problem, family history of cardiovascular disease, obesity, smoking and drinking problem.

Generally, hypertension cannot be treated completely. It could be controlled by two different approaches including the life style modification and using medical treatment. In accordance with JNC report, the ideal level of controlled blood pressure is below 140/90. Some of the common life style changes comprise as losing weight, stop smoking, control alcohol consumption, etc.

The medical treatment of hypertension is usually followed by some distinct groups of drugs that include:

- Diuretics
- Beta-blockers
- ACEIs
- ARBs
- Calcium channel blockers
- Alpha-blockers

- Vasodilators
- Centrally acting alpha-agonists

The following guidelines in Figure 2.1 indicate the process of drug therapy for hypertensive patients (Weber et al. 2014).

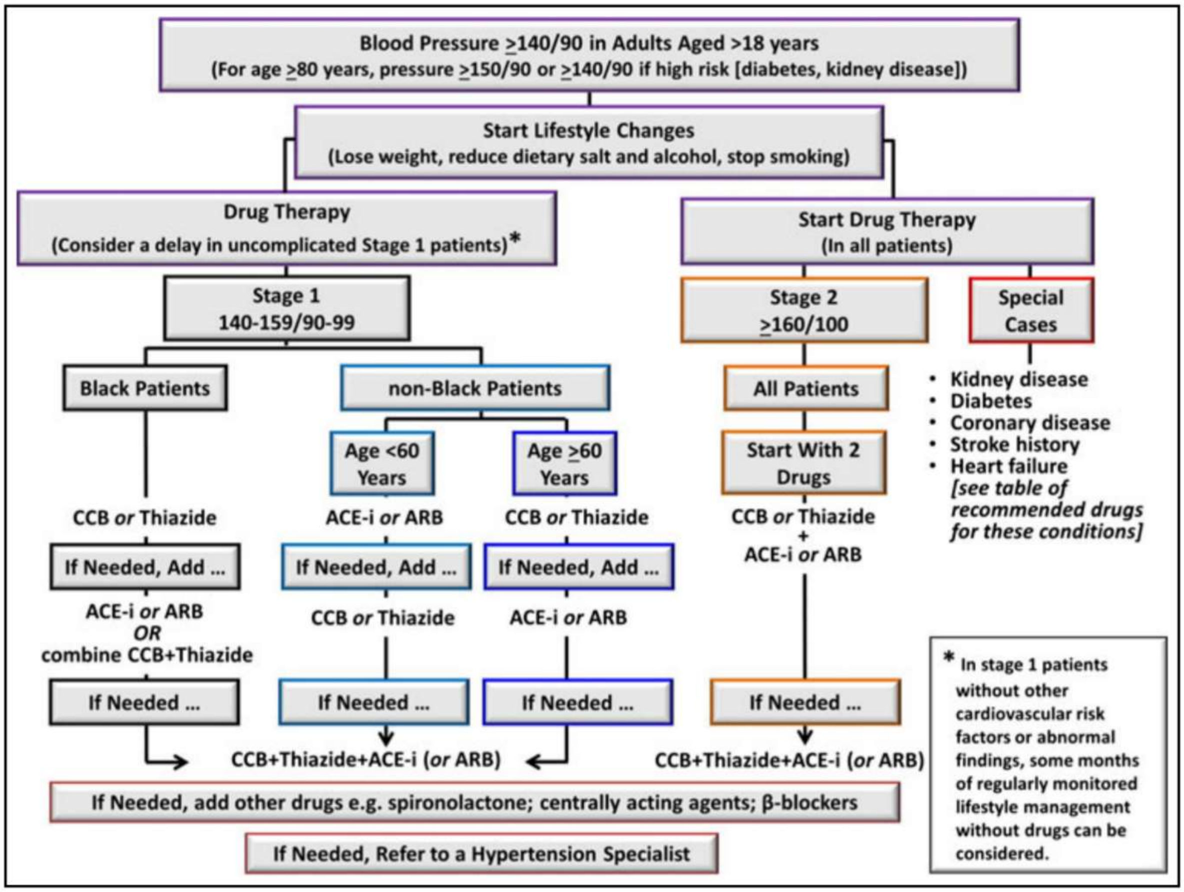

Blood Pressure ≥140/90 in Adults Aged >18 years
(For age ≥80 years, pressure ≥150/90 or ≥140/90 if high risk [diabetes, kidney disease])
Start Lifestyle Changes
(Lose weight, reduce dietary salt and alcohol, stop smoking)
Drug Therapy
(Consider a delay in uncomplicated Stage 1 patients)*
Start Drug Therapy
(In all patients)
Stage 1
140-159/90-99
Stage 2
≥160/100
Special Cases
Black Patients
non-Black Patients
All Patients
Age <60 Years
Age ≥60 Years
Start With 2 Drugs
• Kidney disease
• Diabetes
• Coronary disease
• Stroke history
• Heart failure
[see table of recommended drugs for these conditions]
CCB or Thiazide
ACE-i or ARB
CCB or Thiazide
CCB or Thiazide + ACE-i or ARB
If Needed, Add ...
If Needed, Add ...
If Needed, Add ...
ACE-i or ARB OR combine CCB+Thiazide
CCB or Thiazide
ACE-i or ARB
If Needed ...
If Needed ...
If Needed ...
If Needed ...
CCB+Thiazide+ACE-i (or ARB)
CCB+Thiazide+ACE-i (or ARB)
If Needed, add other drugs e.g. spironolactone; centrally acting agents; β-blockers
If Needed, Refer to a Hypertension Specialist
* In stage 1 patients without other cardiovascular risk factors or abnormal findings, some months of regularly monitored lifestyle management without drugs can be considered.

Figure 2. 1: Hypertensive Medication Guidelines

Despite of the systematic guidelines for treatment of hypertension, the control of this disease hasn’t been quite successful so far. In many different communities more than half of the hypertensive patients are not having controlled blood pressure. The actual cause of hypertension hasn’t been elucidated so far. Many studies were conducted to investigate the effect of genetic and environmental factor on starting the hypertension (Weber, et al. 2014).

As mentioned before, genetic information can define the significant outcome such as susceptibility of the disease that could help to achieve better treatment and control of the disease. The level of hypertension and its related complication such as organ damage, is varied among different patients due to many environmental factors and genetics variations. Therefore, selecting the same treatment approach for all patients wouldn't lead to complete success. Hence, the idea of revolutionizing the hypertension treatment by using genome factors was investigated (Turner, et al. 2007).

The necessity of applying individualized treatment for hypertension is highly recommended by JNC. In 1977, JNC recommended that the initial diagnosis of hypertension should only be related to patient's history and physical examination. However, despite accepting the fact that treatment of patients should be individualized, the first step of treatment for all patients start by prescribing diuretics. Other patients' factors such as demographic information and other comorbidities were used to decide whether drug therapy should be applied or not. Moreover, the effect of prescribed drug on changing the blood pressure wasn't evaluated.

Afterwards, the subsequent JNC reports recommended to apply other patients' parameters like age, race, and comorbidities as the factors for choosing the individualized treatment. However, the mechanism of drugs' action of pathophysiologic changes wasn't assessed (Turner, et al. 2007).

Several years after the first JNC report, Laragh and colleagues (Laragh J.H., et al. 1960) modified a frame work for personalizing the treatment of hypertension by using the renin–angiotensin–aldosterone as an index to effect the blood pressure level. The measurement of plasma renin would define three main subtypes of essential hypertension.

There was no subsequent biochemical measurement, except that plasma renin activity was recorded in order to individualize the antihypertensive drug selection. This could be the result of essential need for extensive knowledge in anatomic, biochemical and physiological mechanisms for blood pressure’s regulation that could lead to better target the malfunctioned organ to lower the blood pressure. Moreover, the expenditure of those measurements was difficult. Therefore, it was decided to perform analysis of DNA cells to extract the genomic variations.

Biomarkers are biological measurements that conduct the most important factors of personalized medicine in three different aspects including diagnosis, risk assessment and predicting the drug therapy result. Therefore, it is believed that applying these measurements could lead to great success to better diagnose and control the treatment of hypertension.

Several studies were performed to determine whether susceptibility genetic variations have intervention with antihypertensive drugs to reduce the risk of coronary heart disease. However, the results demonstrated that there isn’t any distinct cause of phenotype. Since there are multiple genetic factors that can effect hypertensive, they are influenced by environmental factors. Therefore, trying to organize those factors seemed to be more wasteful than helpful. It was estimated that changing the personalized medicine from individual level to larger groups that include more homogeneous patients is a better idea.

2.3 Predictive Algorithms for Personalized Medicine

As mentioned before, genetic information is not a good reference to personalize medicine. There are several genes that contribute to elevate blood pressure and they are dynamically changing while environmental factors are varying.

On the other hand, JNC reports recommended that patients' characteristics such as demographic information, physical examination, comorbidities, etc., can play an important role to better address the treatment of hypertension and reduce the risk of anti-hypertensive drugs side effects. So, in this study, the personalized medicine of hypertension is predicted by using patient parameters.

Predictive algorithms are a powerful tool to help physicians to compare the treatment outcomes of the all patients, analyze the information and predict the best action for individual patients. This information is usually composed to past treatment results or the latest medical research. Prediction models use different techniques such as neural networks to implement the algorithm using the past records of individuals. Then it is applied for new patients for a predictive response.

There are different advantages of using predictive algorithms ("Seven Ways Predictive Analytics Can Improve Healthcare" 2016):

Increase the accuracy of diagnosis: Predictive analysis can help clinicians have more accurate diagnosis judgement. As an example, it is important to decide whether the patients coming to ER need to be hospitalized or not. Efficient predictive algorithms can help the doctors to better evaluate the patients' health status and decide to send the applicants home or hospitalize them.

Moreover, the predictive algorithms can help to identify the possibility of developing some diseases by knowing the genetic information such as early Alzheimer's disease. Therefore, several apps could be installed on patients' phones to track the patients' status in many different aspects such as doing exercise, having healthy diet and other mental activities that is recorded on patients' portal. Therefore, based on those results the health conditions could be predicted and the gene therapy could be initiated as an individualized treatment for patient's specific gene.

Predictive analysis can prevent the appearance of some diseases in public health: The instigation of some diseases could be prevented by using predictive analysis. By knowing some data like genetic information, the risk of diseases could be assessed in advance. Therefore, by changing the life style or other arrangements those complications can be avoided.

Predictive analysis could also help doctors to better individualize the treatment: Using evidence based methods one could direct the right treatment for similar groups of patients rather than individual cases. However, predictive analysis would help clinicians define the exact treatment for individuals and prevent the complications due to drug's side effects.

Predictive analysis can help patients to have better treatment results: In traditional medical treatment procedures patients are prescribed the medications that are used for large group of people. However, predictive analysis can help patients to benefit from the right medication that works for them. Moreover, applying predictive analysis affords important information such as alerting the risk of the disease due to the genetic information or current health condition. That helps patients not only to be more prepared but also collaborate with their physicians to better control the disease by providing the accurate information about their conditions, update their health's portal through the apps and feeling more responsible to observe their life style.

As mentioned before, patients' data is the key part for personalizing the treatment. The data could be comprised of genetic information, clinical data trial or electronic health records. It can identify optimal therapy by analyzing the drug discovery process and also defining the best judgement about patients' health condition by comparing the previous evidences ("Seven Ways Predictive Analytics Can Improve Healthcare" 2016).

In general, predictive analysis comprises different statistical, data-mining and machine-learning algorithms. However, the type of the data and objective of the model can define the most suitable algorithms to be used. As an example, for remarking the different groups of objects the clustering algorithms is usually utilized. For models that need a recommender system, the classification algorithms should be applied. If the model needs to be predicted for the future outcome the regression algorithms is operated.

Data mining method is composed of two models including predictive and descriptive. In predictive models, once the pattern is recognized the future outcomes would be assessed. While in descriptive model, the pattern on the data is used to exploit meaningful outcome like clustering.

In this study, the Case Based Reasoning (CBR) model is used as a descriptive and predictive tool for individualized treatment of hypertension. CBR is a problem solving approach that is originated from cognitive psychology. In this model, the problem-solving model is formalized based on a simple rule of thumb; that is referring the solution of previous problems to arrive at the current problem’s solution. The idea of this model is to store the past problems and adapt the new solution by modifying the past results. Therefore, the history of patients plays an important role for solving the problem

The quality of case based reasoning depends on the following factors:

- The previous cases
- The level of matching the new problem with the previous cases
- Adaptability of a case
- The saving ability

2.4 CBR

Basically, past experience can offer valuable intuition for solving new problems. Case-based reasoning (CBR) is based on the premise that once a problem is solved, it is more efficient to reference it for the next similar case (Janet Kolodner 2014). The ability to understand the new problem by using the old experience has two parts: mining the old experience and interpreting new problems by using the old cases. The more the number of past cases, the richer the reasoner be to achieve the new solution. This technique is based on the two facts of nature. First, similar problems have similar solutions. Second, future problems are usually similar to current problems (Leake 1996). The simplest form of CBR has four steps which focus on composing the interpretation and finding the solution for the new case.

1. Retrieve cases

2. Reuse the knowledge in the retrieved cases to find a solution.
3. Revise (adapt) the similar solution to fit the new problem
4. Retaining the new solution to serve as a reference for the future problems.

The first step of CBR is the indexing and retrieval process. In this step, the features that can be used as significant predictors are identified. Those features could be defined directly by the human expert or they can be selected computationally. Second, the similar cases are identified through the retrieval process. The interpretation of new situation is entered by comparing the old experience with new cases.

After retrieving the similar cases, since there is no exact old similar case to a new one, the solution usually won't be the same. Hence, the solution of previous cases should be modified to better fit the new case. The last two steps are known as the adaptation process which is fixing up an old solution to meet a new problem's demand.

The adaptation could insert a new item into an old solution, remove or revise the current solution. In order to learn from the experience, a human expert needs a feedback to check the reliability of the solution. Therefore, evaluation and repair are the other parts of the case based reasoning.

2.4.1 Case Retrieval

The CBR method uses memory to solve problems. Therefore, suitable solutions from previous problems can be used for new problems. This is based on the relation between the new problem and the previous cases which is known as similarity. Similarity is the process of comparing the cases attribute by attribute. This comparison is based on two concepts including value and weight of the attributes. The similarity is also expressed in different degrees. Therefore, the most similar cases can be chosen as a solution for the new cases.

Different machine learning methods are used to measure the similarity and retrieve the cases. Stottler and Broder (1989) have used Sequential/ non-Sequential indexing method. The number of attributes can define the retrieval time (King, et al. 1992). In 1992, the K-nearest neighborhood was applied for measuring the different degree of similarity (Lewiston, NY 1992). However, this technique has some limitations such as including irrelevant features in the retrieval process that could affect the accuracy and also increase the calculation time. Over the years, different data mining techniques are combined with K-NN to improve the efficiency and accuracy of the results, such as feature selection, feature weights and feature clustering (Krishnaswamy, et al. 2014). Moreover, Rezvan and Hamadani applied rough set theory to eliminate those features and also increase the efficiency and accuracy (Teghi, et al. 2014).

Association rules technique is also a dynamic procedure to extract the best set of rules that can distinguish the different features (Krishnaswamy, et al. 2014). The other disadvantage of K-NN is its non-efficiency for large datasets. Osborne et al. applied different framework for measuring the similarity based on sensibility of cases rather than the problem definition. However, it wasn't useful for adapting the retrieved cases (Osborne, et al. 2005). Afterwards, Mi (2008) applied Grey theory for finding the nearest neighborhood in banking datasets. In this method the weight of each feature was calculated through Analytic Hierarchical Process (AHP). Grey relational theory is a procedure of measuring the level of similarity between two data records which are introduced in a gray system theory. It means the relationship between the system's factors are not defined certainly. So the influence degree between attributes are measured. The procedure of this algorithm is as follows.

- Normalize the data

- Calculate the gray relation analysis of data series,

$$\mathcal{E}i(k) = \frac{\Delta_{min}+\rho\Delta_{max}}{\Delta_i(k)+\rho\Delta_{max}} \qquad (2\text{-}1)$$

where, $\Delta_i(k)$ is representing the difference vector of two data series; Δ_{max} and Δ_{min} are the maximum and minimum of this differences respectively; and the coefficient $\rho = [0,1]$.

- Find the gray relational coefficient.

$$r_i = \frac{1}{m}\sum_{k=1}^{m} \mathcal{E}i(k) \qquad (2\text{-}2)$$

This will show the relation degree between two datasets and *m* is the number of features of the case.

The integration of clustering and similarity measurement was also proposed by Fanoiki (2010). So, the new problem was matched to the cluster of cases that share the same feature values. Along with the integration of clustering and similarity measurement, Hui (2009), proposed Self Organizing Maps (SOM) method for case retrieval. Applying this algorithm, increases the accuracy by visualizing the output clusters. Moreover, other techniques have been applied to optimize the retrieval process. Dalal (2011). has defined the objective function of different feature indexes to investigate the similarity between the new case and stored cases. The weight of features and their local similarity measurements play an important role to increase the efficiency and accuracy. By using this technique, first the similar cases are selected by using k-nearest neighbor and the weight of significant features and then similarity is measured by using the weights of the feature:

$$Similarity(N,K) = \frac{\sum_{i=1}^{n} w_i * sim(f_N, f_{old})}{\sum_{i=1}^{n} w_i} \qquad (2\ \text{-}3)$$

Where f is the feature value of new and old cases. After defining the similarity, the similarity table will be generated to identify the most similar cases. Therefore, the last step is to find the most similar case among the retrieved cases which were defined through local similarity method. Therefore, the global method is used to calculate the similarity value of each feature by using the pervious similarity value and its weight which is calculated as follows.

$$Sim_i = \sum_{i=1}^{n} w_i * sim(N,M) \quad \textit{Where: } \sum_{i=1}^{n} w_i = 1 \qquad (2\text{-}4)$$

Increasing the case attributes could not only could increase the retrieval time enormously, but, it could also degrade the efficiency. Lichcuan et al (2013) proposed the algorithm based on vector model for measuring the similarity. Each case is represented as a vector, so the angle between each vector would define the similarity between cases. Therefore, the weight of each factor can represent the vector model. Moreover, Fernando and Henskenes (2015), have applied orthogonal vectors projection for measuring the similarity of clinical datasets. This technique was covered by cosine similarity, Euclidean distance, and neighbor algorithms. The patient parameters are compiled as a vectors which is projected to some standard vectors for defining the similarity and finally the diagnosis of the disease. Despite the different techniques used for improving the similarity measurement and finding the most relevant cases, there is always a possibility of choosing the wrong match which could cause increasing the decision cost. Hence, Castro and Navarro (2009) have introduced the loss and gain function that could measure the negative and positive consequences of each decision. In order to select the most suitable case and solution for the problem, the probability of occurrence of each solution along with the conditional probability of each solution in accordance with the different attribute value are calculated by using the iterative application of the Bayesian Theorem. Each solution would have the loss and benefit functions.

In addition to the above mentioned methods, Relevance Network has been known to increase the efficiency of the case retrieval. In general, the relevance can be defined in different ways and each definition refers to the performance of different tasks. In Martinez and Campos (2006) study, the number and type of attributes are customized according to the characteristics of the domain. The relevance scales are distributed in six regions from 0 (irrelevant) to 5 (extremely relevant). This approach is a rapid approach to define the feature weighting by replacing the relationships by some rules. The relevance of the attributes is also defined in accordance with the expert domain knowledge that identify the types of relevancies the attributes can have. Therefore, the in-contextual relevance shows the attributes whose values are not affected by other attributes. The in-contextual relevance refers to the attributes whose values are dependent on other attributes. The three important attributes with their degree of relevance are shown in Figure 2.2.

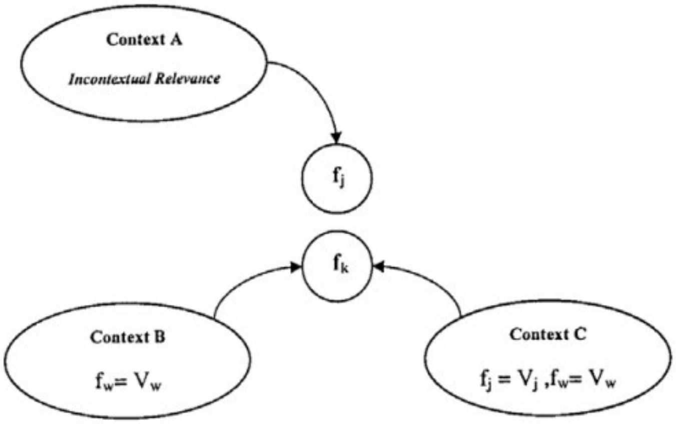


Figure 2. 2: In-Contextual and Contextual Relevance Network

The combination of different problem solving and knowledge representation methods is a very active research area in Artificial Intelligence which is known as hybrid systems. The major effectiveness of a hybrid system is its applicability to various areas which is also fit to solve complex problems (Khandelwal and Sharma 2015). HSU and Ho (Hsu et al 2004) have presented the combination of neural network and fuzzy system for hybrid case retrieval. The input to the case retrieval is the fuzzy patient specification which is represented in two layers. As it can be seen from the Figure 2.3, the first layer shows the symptom net which define the similarity of subjective findings, while the second layer is pathological and laboratory data. Applying the NN and fuzzy logic enables retrieval of the cases which are classified by a decision tree in the next step to calculate the expected utility value.

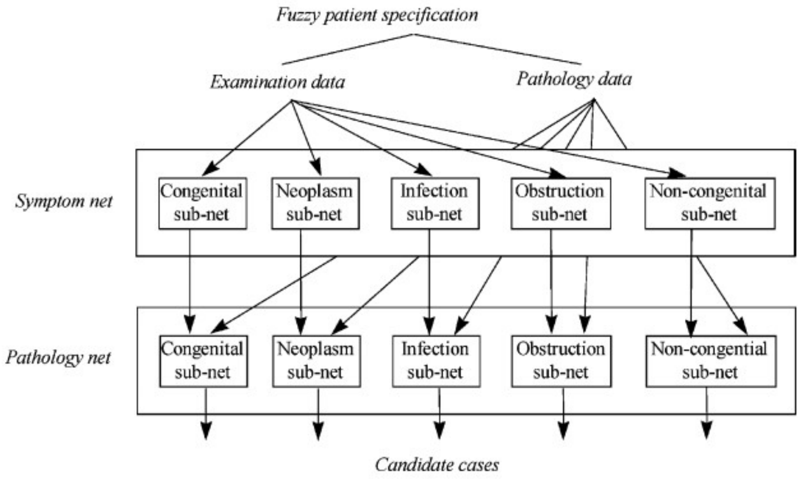


Figure 2. 3: Fuzzy Neural Network

Patients with similar symptoms are being retrieved through the first layer. Those are fed in to the second layer for further investigations.

2.4.2 Case Adaptation

After cases are retrieved, they are adapted to new cases. In general, the reuse and revise steps in CBR is called the adaptation. Adaptation is a necessary step after retrieval. Since, after finding a similar match for the new problem, there is always a possibility that two cases are not identical. Therefore, the selected solution should be revised to match the new case specifications. The complexity of adaption is dependent on the type of the problem. It could be as simple as substitution or as complex as modification the new structure for the solution. In order to be able to adapt a solution, the aspect of a case situation, the reasonability of changes and the control of adaptation process should be considered (Leake, et al. 1995). Therefore, there is no general guideline for the adaptation process. However, the general outline can be defined as follows (Mitra and Basak 2005):

Input:

- A problem description
- An incomplete solution

Output:

- The solution that fit the problem

Method:

- Adjust the non- quite matching solution to meet the solution of the problem

Early CBR systems, have applied the solution of 1-nearest neighbor retrieved case directly as case adaptation solution. This was highly dependent on human judgement (Shahina Begum 2009). Afterwards, the idea of selecting *k* similar cases have been presented by Qi, et al. (2012) to increase the accuracy of solutions. The difference between feature values of the new case and the other cases are being calculated. Then, the availability of a feature is estimated and being transformed to adaptability value by applying induction tree and k-NN mechanism. As it can be seen from Figure 2.4, the feature values between the retrieved and new cases are being compared by using induction tree.

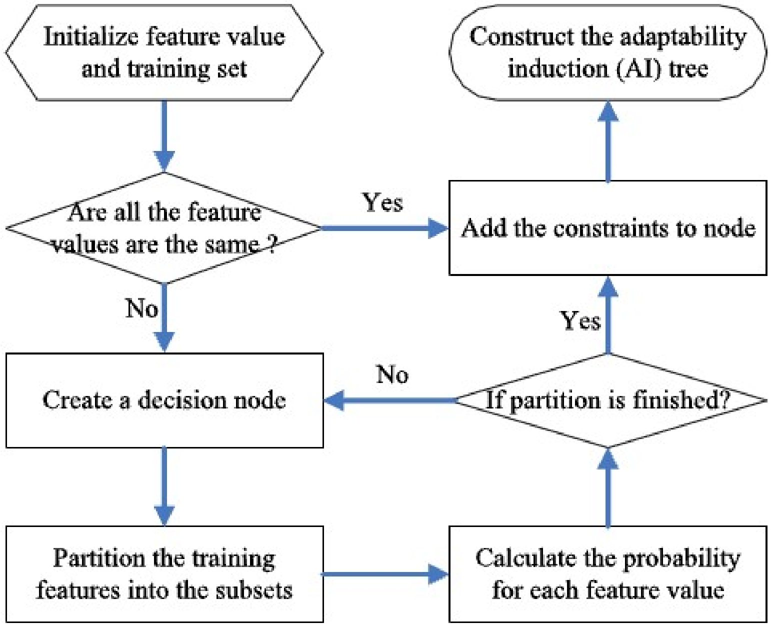

Initialize feature value and training set
Construct the adaptability induction (AI) tree
Are all the feature values are the same ?
Yes
Add the constraints to node
No
Yes
Create a decision node
No
If partition is finished?
Partition the training features into the subsets
Calculate the probability for each feature value

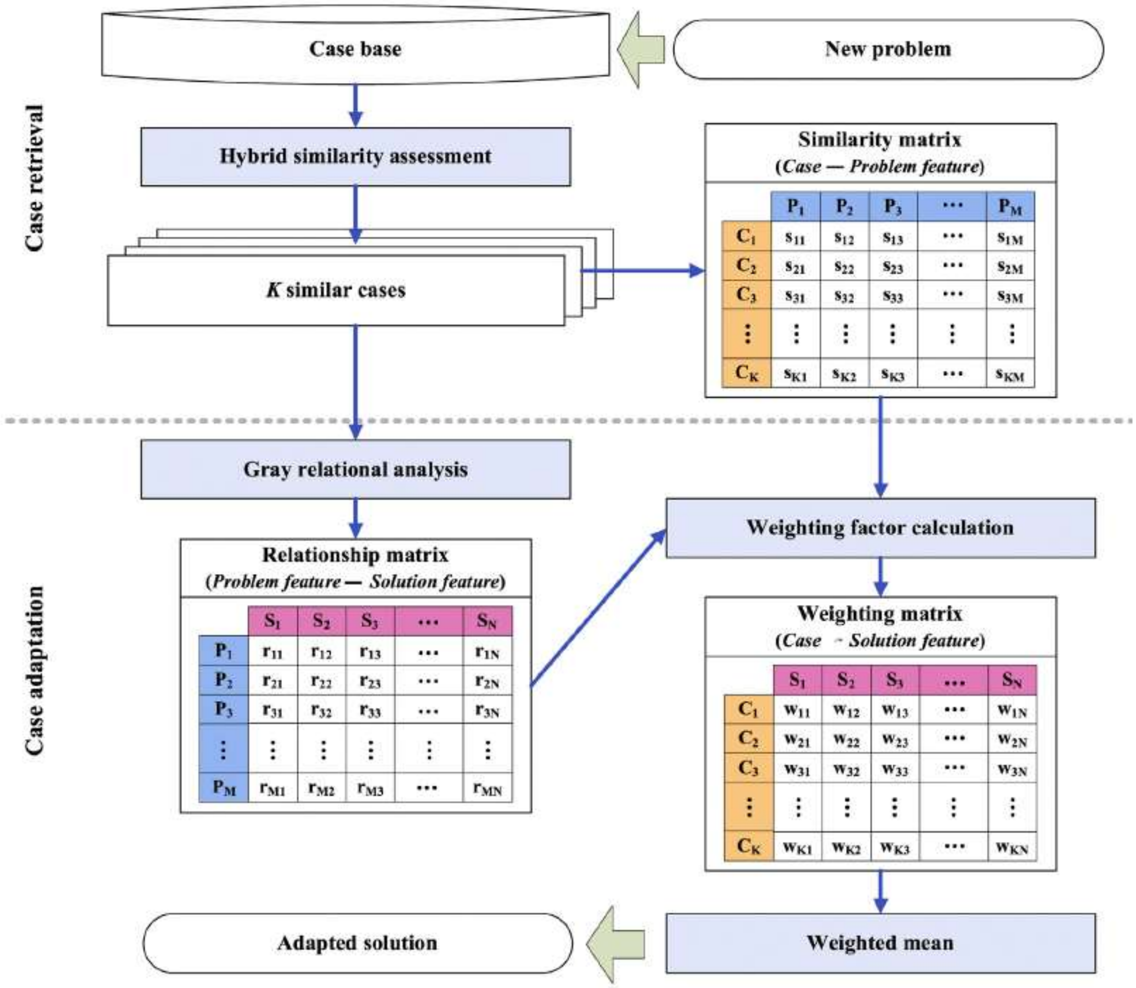
Figure 2. 4: Induction Tree Steps

Despite the specific application of adaptation algorithm, Morello and Haouchine (2013) have proposed a general adaptation model for diagnosis of the fault and repair of industrial equipment. The dependency between case features and class label has been considered. So if the variation of feature could change the class it has been known as high impact. Therefore, three types of relations have been identified: high, low and no relation. Finally, if the solution class of the best chosen retrieved case is similar to the new case class, then the algorithm uses the hierarchical model. If the class is different, then the contextual model would be applied to localize a set of potentially failing components and then uses the hierarchical model. Moreover, Hanny and Keane (1996) applied “case difference heuristic” method for case adaptation. This method is based on difference vector which shows the applicability condition between new problem and similar cases. First the similar cases are being recognized by using k nearest neighborhood algorithm. Then, the adaptation rule is being generated based on the differences. The most important disadvantage of this method is the inconsistency and overlapping of the generated rules.

Different case adaptation methods are preferred statistically or by using intelligent techniques. However, intelligent methods can produce more accurate adaptation results than statistical methods in general. On the other hand, the large amount of data would be needed and the computational cost may also increase significantly (Qi, et al. 2015).

In this regard, different surveys have combined different machine learning methods to get the adaptation knowledge. The advantage of these methods is the independency of the domain knowledge. The utilization of those techniques is to apply the adaptation knowledge enhanced from training data and implement them to automate case adaptation. Along with this idea, Qi, et al. (2015) have used the SVM (Support Vector Machine) algorithm to get the optimum network structure and global solution. The idea of this method was based on regression approximation model to better address the relation between case features and target class. Moreover, they have applied a decision tree algorithm to adapt the case solution. They utilized the induction tree based on different solution features and calculated the adaptability value based on each decision tree (Qi, et al. 2012). The adaptability value is obtained by calculating the difference between new and retrieved cases features which is revised by availability estimation.

As mentioned before, the adaptation process can be done either by using statistical or intelligent techniques. The most common statistical methods which have been designed to include closet analogy method, equal Mean, Median, weighted mean and multivariate regression analysis. To this end, Hu and Peng (2015) have employed the weighted mean algorithm for parametric machinery design (PMD) which is applicable in datasets with small number of records and many parameters (Hu, et al. 2015). The weighting factor of each retrieved solution is calculated by multiplying the similarity and relational matrix which are obtained through weighted average of the solution of the *k* similar cases as shown in Figure 2.5.



	P_1	P_2	P_3	⋯	P_M
C_1	s_{11}	s_{12}	s_{13}	⋯	s_{1M}
C_2	s_{21}	s_{22}	s_{23}	⋯	s_{2M}
C_3	s_{31}	s_{32}	s_{33}	⋯	s_{3M}
⋮	⋮	⋮	⋮	⋮	⋮
C_K	s_{K1}	s_{K2}	s_{K3}	⋯	s_{KM}

	S_1	S_2	S_3	⋯	S_N
P_1	r_{11}	r_{12}	r_{13}	⋯	r_{1N}
P_2	r_{21}	r_{22}	r_{23}	⋯	r_{2N}
P_3	r_{31}	r_{32}	r_{33}	⋯	r_{3N}
⋮	⋮	⋮	⋮	⋮	⋮
P_M	r_{M1}	r_{M2}	r_{M3}	⋯	r_{MN}

	S_1	S_2	S_3	⋯	S_N
C_1	w_{11}	w_{12}	w_{13}	⋯	w_{1N}
C_2	w_{21}	w_{22}	w_{23}	⋯	w_{2N}
C_3	w_{31}	w_{32}	w_{33}	⋯	w_{3N}
⋮	⋮	⋮	⋮	⋮	⋮
C_K	w_{K1}	w_{K2}	w_{K3}	⋯	w_{KN}

Figure 2. 5: CBR Adaptation by Using Similarity and Relational Matrix

Moreover, Li, et al. (2009) have applied the inverse distance weighted mean algorithm to improve the effect of similar cases. The adaptation value will be defined by the cost and similarity value of similar cases.

Besides the statistical and intelligent methods, some other techniques include by applying the hybrid method which implements the combination of these two techniques for adapting a case. As an example, Patterson and Rooney (2002) have utilized the automated localized technique for case adaptation by combining the k-NN algorithm and regression analysis (Patterson, et al. 2002). First the k nearest neighbors are defined. Then by combining these cases based on weights for their individual attributes, the general case is formed. This generates the regression function which predicts the difference in the output attributes between the two cases. Moreover, Jung and Lim (Jung, et al. 2009) utilized hybrid approach by combining the neural network and *k*-clustering methods. First the cases are being clustered by using k-cluster technique. Then, the adaptation knowledge is extracted by employing the Radius Basis Function (RBF). In such a way, the center of the clusters is being modified by recalculating the weight between the output layer and the new design layer. However, this method is applicable just for numerical data.

The premise in CBR system is to generate consistent solution with domain knowledge. However, there are some solutions which are inconsistent with expert knowledge and known as a failure. Basically, it is impossible to completely remove the gap between expert and domain knowledge in CBR systems which is called qualification problem. Cordier and Fuchs (2007) have grouped these failures to partial and non-partial. After the solution of the target problem has been defined, the expert knowledge will be applied to revise the domain knowledge and fix the problem. They applied this method for treatment of breast cancer. Therefore, the similar cases will receive the same treatment such as chemotherapy and radiotherapy. However, for the male patients this treatment is not partially correct since men don’t have ovary. The repairing algorithm will fix this inconsistency.

2.5 CBR Applications

CBR has many applications and has been used in different areas. One of the application is in medicine as a medical support system and healthcare planning (Huang, et al. 2007). Applying CBR is growing rapidly in healthcare area in both research and practice. There are many CBR-based systems used as medical decision systems. The major ones are CASEY, NIMON, ICONS and FLORENCE. CASEY deals with diagnosis of heart failure; NIMON is used to monitor the renal function; ICONS presents antibiotics therapy for treatment of patients (Koton 1988), (Wenkebach 1992), (Schmidt and Gierl 2001); and FLORENCE system deals with nursing planning (Welter et al. 2011).

2.6 Temporal CBR

In some area of CBR, the problems are continuous in the time domain. It means that the new records are produced dynamically or retrieving the current case isn't just dependent on individual case but on stream of cases' history. The standard CBR is not useful. Therefore, in these cases the temporal CBR is applied. This approach can be used for dynamic supervision of events. In the medical field this can be applied to diagnosis, prognosis or tracking the patients' health status. Sànchez-Marrè et al. (2005) have applied a new method in temporal CBR by using episode-based reasoning. Each stream of a case is shortened into episodes. In general, applying episode-based CBR might develop some challenges such as formation of the episodes, dynamically increase of their length, similarity evaluation and their retrieval. However, in Sànchez-Marrè et al. (2005) study, those challenges were addressed properly. Each case has been defined by following structure in library:

C_t = <CI, t, CD, CDL, CS, CE>

where, CI is case ID; *t* refers to the time of the records; CD shows the information of other attributes; CDL is diagnosis result; CS the solutions that have been used and CE would be the evaluation of the solution. In order to create the episode-based CBR on the cases the following structure is generated:

$E^d_{t,l}$= <EI, t, l, ED, d, ES, EE, C_t, C_{t+l-1}>

Where EI refers to the episode ID; *t* shows the initial time; *l* is the length of the episode; ED is the episode description; *d* episode diagnosis and ES is the episode solution; and EE is the evaluation of episode. The C_t and C_{t+l-1} represent the start and end case for each episode. The retrieval process is accomplished by using discrimination tree that defines which episode would be the best match. However, in this approach different episodes might share the same cases. Figure 2.6 is showing the retrieval process of each episodes.

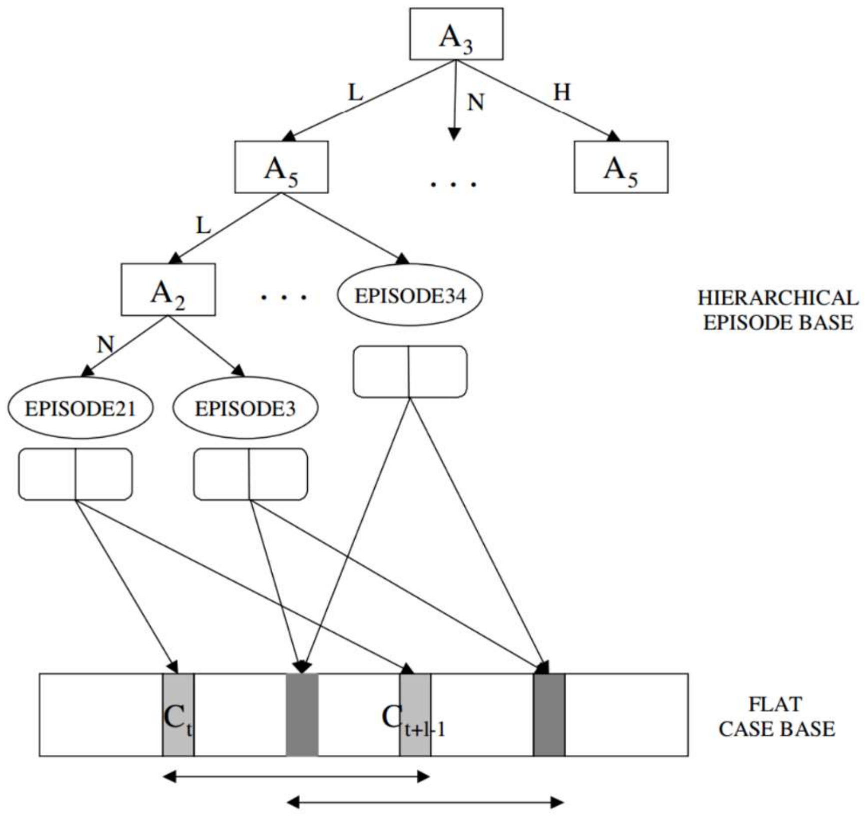


Figure 2. 6: Discrimination Tree for Episode Retrieval

In Hafiz and Hassin study, the applicability of using temporal CBR for controlling the reservoir spillway gate has been considered levels (Hassin, et al. 2006). The decision making system for opening and closing the gate based on the water level was created. Moreover, this system could be helpful for new engineers to develop their skills by using those experiences. The hydrologic data are in the form of time series. So, in order to retrieve the time series data, they are segmented based on different time slots. Since the process of gate opening is dependent on water level and the level of water changes after rainfall, there is always a delay. In this state a window is to capture that delay. The operation of the gate is tracked over different time period and water.

The other application of temporal CBR has been done in hemodialysis area. In a hemodialysis procedure, the treatment process is given three times a week for four hours. This information is usually stored as time series abstraction (TA) data. Retrieving this information would help physicians to better visually analyze the results of treatment. The information of each TA data is composed of patient characteristics and the case solution. The solution would be in the form of time series signals that are composed into two important elements: state and trend. In the retrieval process similar cases would have the optimized similar combination of state and trend. During the retrieval procedure, those signals are being processed by a TA processing module and the output would categorize the state, trend and their combination. However, this approach could have some complications such as missing abstraction and low quality input (Montani et al. 2009).

Chapter 3: Methodology

3.1 Introduction

The focus of this research is to better address the medical treatment for hypertensive patients based on their personal characteristics, this applies the concept of personalized medicine to hypertension treatment Different hypertension risk factors are used as indicators to better tackle the stage of the disease and finding the best treatment.

To this aim, the temporal Case Based Reasoning technique (CBR) is applied to:

- Retrieve the most similar patients as a group
- Reuse the previous treatment from the most similar patients
- Revise their treatment to meet a new patient's case
- Retain the proposed solution in case base for further

In this regard, the medical records of patients enrolled at Massachusetts General Hospital (MGH) and affiliated with Harvard Medical School are analyzed.

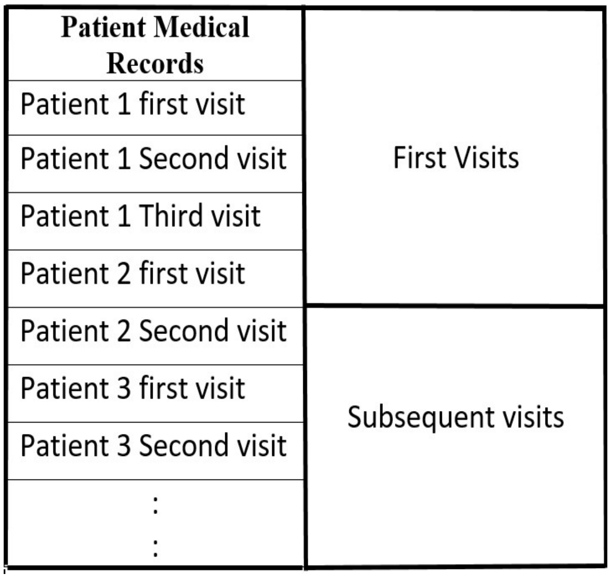
The data is provided in different batches. Each batch consists of different files that cover patient demographic, diagnosis, lab results, physical examination and medication of diabetes patients who were also suffering from other comorbidities like hypertension.

3.2 Proposed Temporal CBR for Hypertensive Patients

For the treatment of hypertension, the history of patients plays an important role to define the best medical decision. Since, patients have different response to hypertensive drugs or they might have different side effects to drugs, the treatment regime might change due to those reasons after each visit until the SBP reaches the optimum level.

Hence, the visit history of patients should be reflected in the analysis.

Based on previous studies, different types of temporal CBR have been applied for handling the time based problems. For some of those problems, the start and ending the process has definite length and the information was captured at equal time interval, such as drug injection for hemodialysis patients. So, using the temporal CBR would be straighter forward. Moreover, in episode based reasoning, the whole history of patient would be defined as an episode. However, this approach could dynamically increase or overlap with other episodes that wouldn't meet the requirement of case retrieval of this problem, since the patient's status should be tracked over each visit to find the better path of treatment. In this study, patients have different number of visits during their treatment process and the time intervals of their visits were not the same. So, in order to solve the problem, a new approach of temporal CBR has been applied. The concept definition of cross sectional analysis has been used to capture the records of first treatment visits of patients. Considering the fact that patients have no previous medical history at their first visit, we remove the history issue and turn the similarity measurement into the simple form. All patients have started their treatment after first visit and by comparing their characteristics, the similar cases would be identified. This procedure is shown in Figure 3.1.



Patient Medical Records	
Patient 1 first visit	First Visits
Patient 1 Second visit	
Patient 1 Third visit	
Patient 2 first visit	
Patient 2 Second visit	Subsequent visits
Patient 3 first visit	
Patient 3 Second visit	
: :	

Figure 3. 1: Dividing Patients' Records into First and Subsequent Visits

After dividing the dataset into two groups of First and Subsequent visits, the retrieve phase of CBR has been applied as follow.

3.3 Retrieval of Temporal CBR

The first step of CBR is to identify the most similar cases from the case base that is called the retrieval process. This process compares the new case's parameters with all previous cases and select the most similar candidates as a reference for subsequent steps. Since this process is repeated for each new incoming patient, there should be a structured formation of similar groups of patients in the case base. To this aim, a clustering algorithm is used to group the patients' cases with similar characteristics.

To this aim, the k-means cluster algorithm, has been applied to define the cluster group of each patient.

k-means clustering is one of the most common partitioning method. In this method, first different number of clusters "K" are defined. Each k represents the centroid of a cluster.

Then for each centroid, the closest data points are assigned. This process iterates until the centroid of each cluster reaches the average of all data points and the sum of square errors reaches the minimum level.

Selecting the similar cases at the start point of treatment would indicate that the treatment path could be similar for matched patients. Therefore, in this problem the case retrieval process is proposed through the similarity measurement of first visit dataset. The process of finding the similar matches and extracting their subsequent records is shown in Figure 3.2.

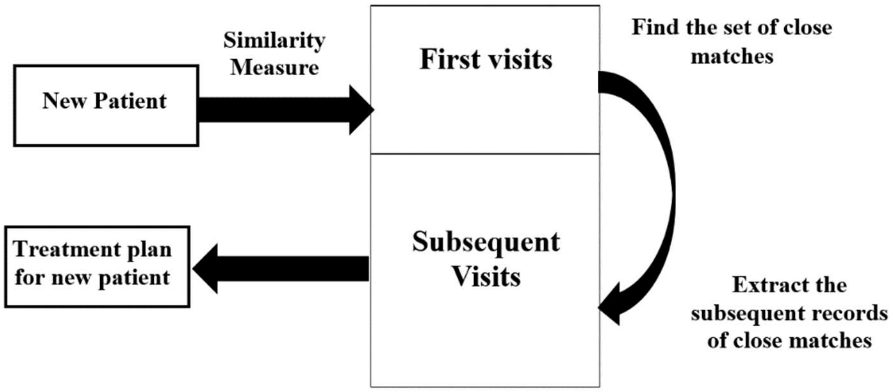

Similarity Measure
New Patient
First visits
Find the set of close matches
Treatment plan for new patient
Subsequent Visits
Extract the subsequent records of close matches

Figure 3. 2: CBR Retrieving of Similar Matches

3.4 Adaptation of Temporal CBR

After retrieving the similar cases, their results are adapted to meet the new case's requirements. Based on different similarity results, different adaptation approaches are chosen to address the solution. The adaptation phase is usually composed of two steps that is called reuse and revise. In reuse step, the proposed solution of similar cases is obtained and checked to meet the required solution for the new case. Sometimes it can be directly applied to the new case and sometimes they need to be changed and meet the revise step. In this study, both steps have been defined and applied to meet the new patient's treatment.

3.4.1 Reuse of Temporal CBR

After clustering the similar groups of patients, their follow-up visits are extracted from the Subsequent Visit dataset. In order to reuse this information, the result of each follow up is reviewed at each visit. If the patient's blood pressure has been dropped to lower level or stabled at the lowest level at the next visit, the previous treatment is considered successful successful.

Based on the fact that similar patients should be given similar medications, we hypothesize that the successful results of similar patients would give the successful result for the new patient as well. Two different perspectives are studied to collect more details about the successful results. Those are successful visits and successful treatment process. Figure 3.3 shows the process of selecting successful results based on subsequent records of similar matches.

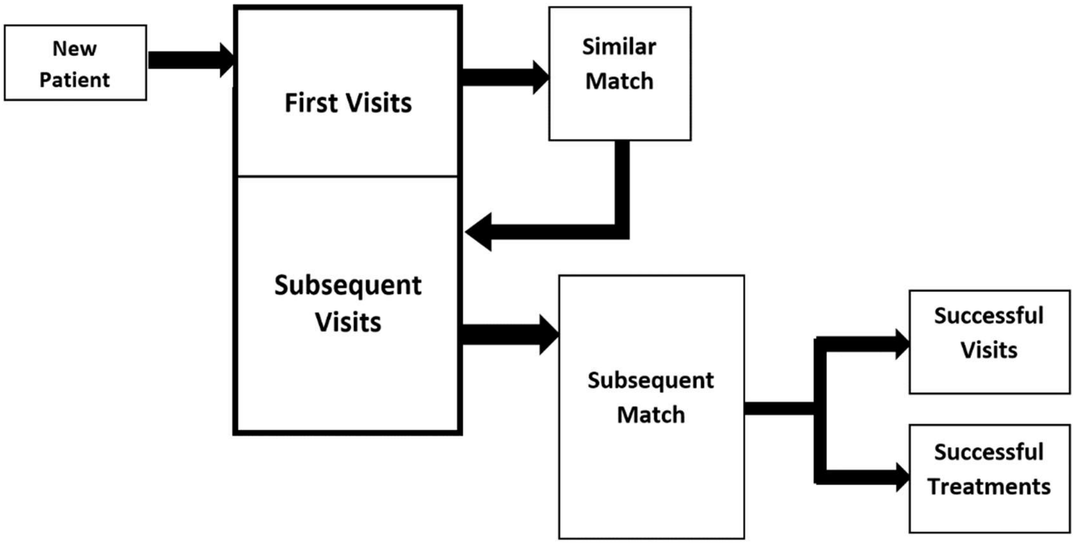


Figure 3. 3: Successful Results of Subsequent Matches

3.4.2 Revise the Temporal CBR

In the revise step of CBR, the treatment procedure of each similar candidate is modified to meet the best result. This can be done through successful visit and successful treatment.

Successful Visits

A successful visit is a visit *i* whose prescribed drugs could lower the blood pressure or stabilize the blood pressure at the lowest level at visit $i+1$. In order to find the successful visit, the preceding records of similar patients will be required. So after selecting the best matches, this information could be extracted from the Subsequent visit dataset. The episode model will be generated for each two following visits. The information of each episode consists of blood pressure level at visit *i* and $i+1$, the medication at visit *i*, and the patient ID number. Therefore, the successful visit will be defined by realizing the difference between blood pressures during two consequent visits. Figure 3.4 shows the process of selecting a successful visit. Each graph shows the changes of SBP level for each individual patient during their treatment process.

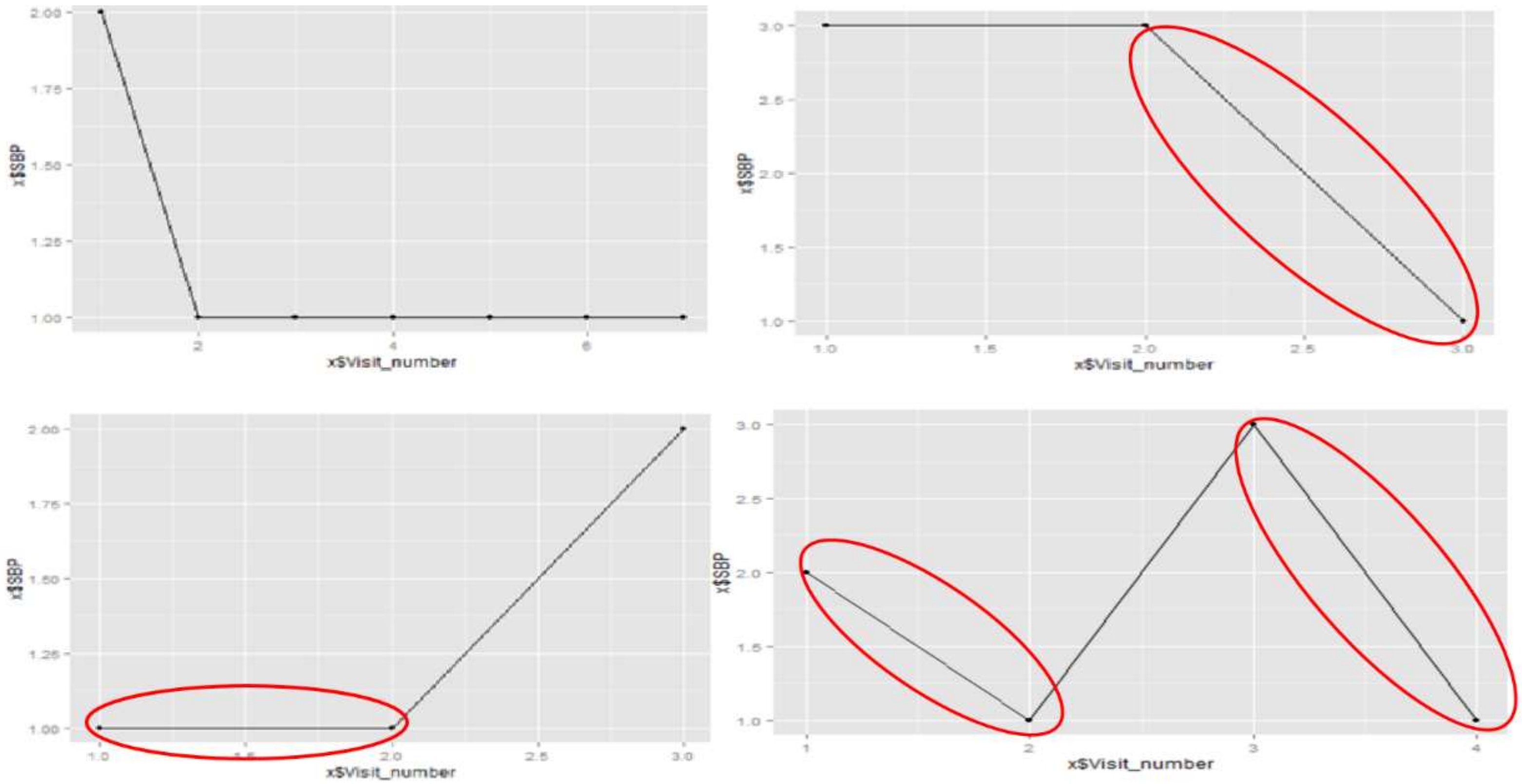


Figure 3. 4: Selecting the Successful Treatment for Each Patient During Visits

As it can be seen from the Figure 3.4, each graph shows the SBP level changes of individual patients over number of visits. For each graph, the successful treatment at visit *i* that could reduce the SBP for the visit *i*+1 has been circled. As an example, for the graph in upper left, all the visits were successful, since they reduce the blood pressure and stabilized at the lowest level. So after selecting the successful visits, the new dataset will be created as shown in Table 3.1.

Patient ID	SBP *i*	SBP *i+1*	Medication *i*	Difference
100	3	2	A+B	1
101	2	1	A	1
102	1	1	C	0
:	:	:	:	:

Table 3. 1:Treatment Successful Dataset

The patient ID and the date of visit *i* would be used to extract the other information of patients from the main dataset. The new dataset will indicate the successful visits of similar cases for the new patient.

Using successful visit information would help physicians to find out some important information about the impact of different medication regimes to stabilize or lower the blood pressure. To this aim, the SBP level at visit *i* would indicate the impact of medications. If the SBP level at visit *i* is greater than 1, it means that the prescribed medications have lowering effect. SBP level equal to 1 means the prescribed medications are stabling the level.

Moreover, the successful visit approach could also be used to predict the most frequent medication for the new case. Therefore, using classification algorithm would help to find the predictive model of each drug under patient's different characteristics. In this study, nine different groups of medication are used as an output label. Therefore, in order to classify the treatment result, the common classification algorithms such as Decision tree won't be helpful. To this aim, the multi-label classification algorithm is applied to categorized the results of different medication.

Multi-label classification is a powerful tool used in different new technologies such as categorizing the text, music, movie or in different healthcare area like audio event classification (Peng et al. 2009). In this study, the multi-label classification has been used to analyze the multi-label dataset. The average number of medication for each patient, the imbalance ratio of each drug and the frequency of each drug are defined through this method.

Different multi-label algorithms exist like adaptation, transformation, binary relevance (BR) and label power set (LP) (Peng et al. 2009). However, the BR and LP are the direct method to multi-label classification. Generally, multi-label classification has two categories, including label-based and example-based. In label-based approach, the classification is defined for each label while in example-based, it is defined for each instances and then get averaged. Moreover, the output of classifier could be in the form of binary output or a ranking value.

The metrics obtained by those methods are defining as follows:

Accuracy

Area Under Curve (AUC)

Mean of multi-label classification error (mmce)

In this study, the “mldr” package in R has been used to apply the multi-label classification. Those metrics are used for measuring the classification performance. The accuracy is the proportion of correctly predicted labels while the “mmce” is the error rate of the classification. AUC defines the trade-off between the true positive rate and false positive rate. Moreover, the “getPredictionProbabilities” function in mldr library, would show the predicted probability of each medication that could be prescribed for the new case. So, when the new patient comes, the physician would have better perspective to find the highly recommended medication that have had good outcome.

Successful Treatments

Besides the successful visits, successful treatments can also be used to better address the treatment. In successful treatment approach, the whole process of treatment for each similar patient is investigated and the procedures that could help to achieve the lowest blood pressure and stabilize it for the last two subsequent visits are selected. The following metrics can be obtained from successful treatments.

- The impact of antihypertensive medication to control the blood pressure

○ That is referring to the ratio of successful treatment to the total number of selected treatments

- The average time of the treatment process and the shortest treatment path

3.5 Applying Temporal CBR for New Patient

After defining the different medication procedure for each group of similar patients, the personalized treatment approach for new patient can be used by comparing the new patient's characteristics with all the cases stored in First Visit dataset. Then, the result would have different options:

- All similar candidates belong to the same cluster
- Similar candidates belong to different clusters

In a situation where all similar candidates have the same cluster, the assigned treatment procedure for that cluster is directly applied to the new patient. However, if the similar candidates don't belong to the same cluster, it means that their clusters have some overlap that is more fitted to the new patient's characteristics. Therefore, the subscription of assigned treatment of each cluster should be applied for the new case.

3.6 Retain of Temporal CBR

After revising the solution for each similar cluster in case base, it needs to be confirmed by expert and then applied to the patient. Then, the record is saved in the case base for the future process.

Chapter 4: Implementation

4.1 Introduction

In this research, the data are cleaned, analyzed and classified using R Studio commercial package. R is a programming language that is a powerful tool for statistical analysis and data mining. There are various statistical and graphical techniques embedded in R libraries that help scientists to better model, analyze, cluster, classify and visualize the data. Different functions in R packages facilitate these procedures.

In this study, different R packages and libraries are installed for each step of processing and analyzing the data and implementing different algorithms.

The first step of this process is data cleaning. The important features are selected and then, the different approached are used to clean the dataset.

4.2 Data Screening

The clinical dataset is composed of different patients' information such as demographic, physical examination and lab tests. In order to use this information, a process should be used to clean, characterizing and filling out the missing values to build the final dataset for further analysis.

The demographic file has the following attributes:

- EMPI
- Gender
- Date of birth
- Age
- Race

- Marital status
- Date of death

However, after the cleaning process, the patients who died, were removed from the list.

The primary disease of the patients included in the data files is diabetes and hypertension is a secondary disease. After cleaning the diagnosis file, only the hypertensive patients were selected from the file. After selecting the hypertensive patients in each batch, their further information has been pulled out from other files.

The diagnosis file is including the following information:

- EMPI
- MRN type
- MRN
- Date
- Diagnosis Name
- Code type
- Code
- Provider
- Clinic
- Hospital
- Inpatient
- Encounter number

Where the diagnosis attribute is defining the different types of diagnosis such as hyperthyroidism, diabetes, abdominal pain, hypertensive renal disease, benign hypertension and malignant hypertension. The selected attributes are as follows:

- EMPI
- Date
- Diagnosis Name

In order to clean the diagnosis file, the hypertensive complications have been selected as Diagnosis name.

The lab file, stores some vaccination information, chest exam, smoking status and different blood and urine test. After studying the different types of tests, the most related parameters that can be considered as hypertension risk factors were selected, e.g. urine test and blood test. The selected elements of urine test can infer the functionality of kidney. High blood pressure could damage the vessels in kidney and cause malfunctions. Thus, tracking those parameters could indicate the state of blood pressure. In this research the Creatinine and Micro albumin elements of urine test were selected as indicator of kidney operation. Some blood factors on the other hand, will help to better evaluate the cause of hypertension, such as blood lipids test, including LDL (low density lipoprotein), HDL (high density lipoprotein), Triglycerides and total cholesterol. When the amount of those factors like LDL increases in blood, it could be accumulated inside the vessels and build a wall that makes circulation difficult and could lead to hypertension and heart disease. Triglycerides on the other hand, is a risk factor of hypertension. The excessive amount of this triglycerides could harden the arteries that cause the loss of the elasticity of blood vessels, leading to elevate blood pressure (Grey 2016). Diabetes is one of the risk factors of hypertension. Because high blood glucose damages the arteries that cause atherosclerosis. Atherosclerosis can elevate the blood pressure and in severe case could lead to blood vessel damage, stroke, heart failure, heart attack and kidney failure (WebMD 2016). Therefore, in addition to blood lipids, blood glucose level is used as parameter to define the hypertension. In data files, the lab result document has the following attributes:

- EMPI
- Date
- Test Description
- Result

Besides lab test, some physical examination could help to better diagnose hypertension. Since patients can have hypertension without any significant symptoms, a careful physical examination can help to better evaluate the patient's health status. Those tests comprise as pulse rate, respiratory rate and BMI (body mass index). However, those tests can only partially define the patients' health status. There are other factors that can affect the blood pressure level such as life style, exercise, smoking and alcohol consumption and stress level. However, due to the lack of sufficient information, those factors were not considered in this study.

After selecting the significant predictors, the records are cleaned up and merged to compose an inclusive dataset that covers all the predictors and the records of different patients. In each record, the patient's ID and the date of their visit is also documented. However, the patients' ID is a confidential attribute and the real values won't be shown in rest of this study. Thus, the whole dataset would cover the history visits of different patients.

The physical examination file information is composed of following attributes:

- EMPI
- MRN
- MRN type
- Date
- Concept name
- Code
- Code type
- Result
- Units
- Provider
- Clinic
- Hospital
- Inpatient
- Encounter number

EMPI (enterprise medical patient index) is used as an ID of patients through all different healthcare systems. MRN (Medical record number) is unique patient ID number that is established separately by each healthcare system. The MRN type would define the type of the hospital or healthcare provider. Concept name attribute is consisting of different types of physical examination of patients such as pulse rate, weight, height, blood glucose, SBP/

DBP, temperature, respiratory rate, pain level, influenza vaccine, smoking and alcohol status. The code and code type is indicating the LMR (Longitudinal Medical Record) code number. Moreover, the result and unit attributes are showing the physical examination and the unit of measurement respectively. The provider, clinic and hospital would also indicate the name of the healthcare system that provided the medical service. The inpatient record is giving the information about the patient situation whether he/she is inpatient or outpatient and finally the encounter number is giving the information about the patients' visit number. However, after reviewing each attributes, the following parameters are selected as the important features:

- EMPI
- Date
- Concept name
- Result
- Units

The Concept name attribute is a column that include different types of physical examination that were applied to diabetes patients. This information was not uniformly provided for all the patients. Hence, during the selection of those parameters, some of them were removed from the dataset and irrelevancy to hypertension problem. The final selected parameters were defined as Pulse rate, Respiratory rate, SBP/DBP, Height and Weight. The height and weight of patients were converted to BMI (Body Mass Index) parameter to better define the state of obesity of patients. The other information such as temperature, smoking status, alcohol consumption and pain level was not efficiently reported to be used as a predictor and was removed from the dataset.

Besides demographic, physical examination and lab results' files, the medication file is also composing of the following attributes:

- EMPI
- Medication date
- Medication
- Quantity

After cleaning the data, nine different groups of medication were selected to address the hypertension treatment. They are Angiotensin converting enzyme inhibitors (ACE inhibitor), Alfa blocker, Angiotensin Receptor Blockers (ARB), Beta blocker, Calcium channel blocker, Diuretics, Antidiabetics, Cholesterol medication and statin.

The first five groups of medications are known as the hypertension drug's classes and are usually applied to reduce the blood pressure. However, each group of medication has a different function. Diuretics are used to reduce the sodium and fluid in the blood by increasing the urination. This group of medication is usually prescribed alone or in combination with other anti-hypertensive medication. Beta blockers reduce the blood pressure by lowering the heart rate. ACE inhibitors decrease the angiotensin production. Angiotensin is the hormone that narrow the blood vessel. So, reducing this hormone will lower the blood pressure. ARB medications has a similar function. They prevent binding the angiotensin to other receptors of the blood vessel. Alfa blockers on the other hand are used to dilate the blood vessels. Calcium channel blocker is another group of hypertension medication that blocks the calcium entry into muscle tissue. Because calcium is a mineral that increases the contraction of blood vessels and heart, it can increase the blood pressure (Mayo Clinic 2016).

Besides the hypertensive medication, other groups of drugs were prescribed in this study. The Antidiabetics is used to reduce the blood glucose. Since, the primary disease of patients was diabetics, antidiabetic medication is one of the common drugs that was prescribed along with hypertension medication. Moreover, the blood lipid could also initiate some heart disease problem that is a major risk factor of hypertension. So, the cholesterol medication and statin are two other major groups of lipid-lowering medications that has been applied. So, for each record, one or the combination of drugs are prescribed.

4.3 Data Clean up

After selecting the significant features, the following steps are used to clean the data:

- Change the metrics of some attributes like weight, height and SBP/DBP level to the same unit.

- Rearrange the table to have different attributes as columns instead of rows.

- Combine some features to form a new metric. E.g. using weight and height to add BMI (Body Mass Index). The following formula is used to convert the weight and height to the BMI:

$$BMI = 703 \times (Weight\ in\ Pounds) / (Height\ in\ inches)2 \tag{4-1}$$

- Change the metrics of some features to have the same unit

4.4 Categorizing the Data

After cleaning the dataset, the attributes are categorized for further analysis. The following tables are shown the different categorical levels of each attribute.

Lab tests	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Cholesterol	$x<200$	$200\leq x<239$	$x\geq 239$	-----
HDL	$x<50$	$50\leq x<60$	$x\geq 60$	-----
LDL	$x<100$	$100\leq x<129$	$129\leq x<159$	$x\geq 159$
Triglycerides	$x<140$	$140\leq x<160$	$x\geq 160$	-----
HbA1C	$x<7$	$7\leq x<9$	$x\geq 9$	-----
Micro albumin	$x<30$	$30\leq x<300$	$x\geq 300$	-----
Creatinine	$x<1.1$	$x>1.1$	-----	-----

Table 4. 1: Categorical Lab Results

BMI	**Levels**
$x<18.5$	1
$18.5<x<24.9$	2
$25<x<29.9$	3
$x>30$	4

Table 4. 2:Categorical BMI Levels

Pulse Rate	**Levels**
$28<x<55$	1
$55\leq x<75$	2
$x\geq 75$	3

Table 4. 3:Categorical Pulse Rate Levels

Respiratory Rate	**Levels**
$x<20$	1
$x>20$	2

Table 4. 4: Categorical Respiratory Rate Levels

Age	**Levels**
20-30	2
30-40	3
40-50	4
50-60	5
60-70	6
70-80	7
80-90	8
90-100	9

Table 4. 5:Categorical Age Levels

Race	**Levels**

White	1
Black	2
Asia	3
Hispanic	4
Other	5

Table 4. 6: Categorical Race Levels

Marital Status	**Levels**
Married	1
Divorced	2
Single	3
Widow	4
other	5

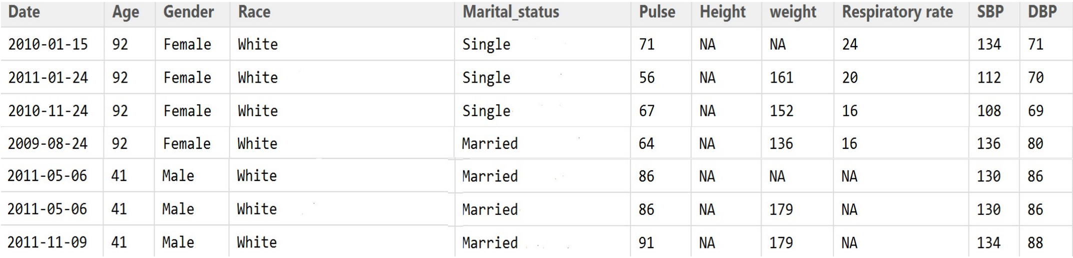
Table 4. 7: Categorical Marital Status Levels

Gender	**Levels**
Female	1
Male	2

Table 4. 8: Categorical Gender Levels

4.5 Filling the Missing Value

Despite of cleaning and categorizing the data, they are not completely suitable for analyzing the process yet since some of the selected attributes have missing values. Table 4.9 is showing the sample of cleaned dataset table.



Date	Age	Gender	Race	Marital_status	Pulse	Height	weight	Respiratory rate	SBP	DBP
2010-01-15	92	Female	White	Single	71	NA	NA	24	134	71
2011-01-24	92	Female	White	Single	56	NA	161	20	112	70
2010-11-24	92	Female	White	Single	67	NA	152	16	108	69
2009-08-24	92	Female	White	Married	64	NA	136	16	136	80
2011-05-06	41	Male	White	Married	86	NA	NA	NA	130	86
2011-05-06	41	Male	White	Married	86	NA	179	NA	130	86
2011-11-09	41	Male	White	Married	91	NA	179	NA	134	88

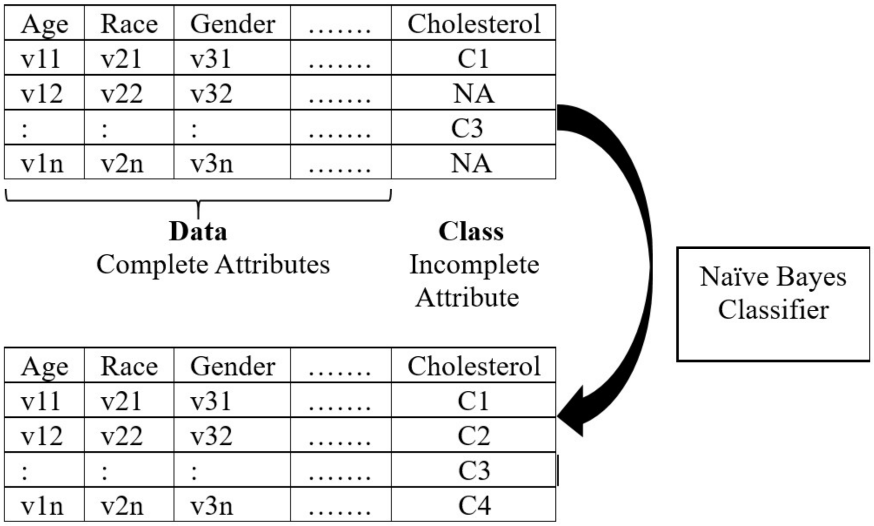
Table 4. 9: Clinical Records Having Missing Value

There are different approaches to deal with missing data, like removing the incomplete records or filling them using different predictive algorithms. The simplest way to fill the missing data is to replace them with the mean of each variable. Moreover, the linear regression and K-nearest neighbor method could also be applied. In regression method, the linear relationship between the variables is assumed to fill out the missing data. However, in most cases these relationships are not linear so the model wouldn't be fitted.

On the other hand, in K-nearest neighborhood method, no predictive model is required for each variable. However, the algorithm wouldn't be so efficient since for each missing value the whole dataset would be searched and that would be a time consuming approach for a large dataset. Moreover, the value of K could change the model's accuracy significantly.

Naïve Bayesian classifier is one of the methods that works very well in dealing with missing data. It is more efficient, simple and accurate that use all set of data to fill out the missing values. The whole variables of the dataset are divided in to two groups, the attributes with having missing values which are called class attributes and the attributes without having missing data that are called training attributes. For each class attribute the training attributes are used to predict the missing values.

In this study, the demographic information, diagnosis type and blood pressure were used as training attributes. Other features like lab results and vital signs are classified as class attributes. Using Naive Bayesian classifier, we filled the missing data of each attributes. Each class attribute is filled by using the training attributes. This process is done through different iterations and has been shown in Figure 4.1(Umamaheswari, et al. 2016).



Age	Race	Gender	…….	Cholesterol
v11	v21	v31	…….	C1
v12	v22	v32	…….	NA
:	:	:	…….	C3
v1n	v2n	v3n	…….	NA

Age	Race	Gender	…….	Cholesterol
v11	v21	v31	…….	C1
v12	v22	v32	…….	C2
:	:	:	…….	C3
v1n	v2n	v3n	…….	C4

Figure 4. 1: Filling the Missing Value Using Naïve Bayes Classifier

The following formula is used to find the missing value.

$$P(class_j|x) = \frac{P(x|class_j) \times P(class_j)}{P(x)} \tag{4-1}$$

By assuming the independency between predictors, $P(x|class_j)$ is equal to:

$$P\left(x\middle|class_j\right) = P\left(x_1\middle|class_j\right) \times P\left(x_2\middle|class_j\right) \times P\left(x_3\middle|class_j\right) \times \ldots . \times P\left(x_k\middle|class_j\right) \quad (4\text{-}2)$$

The $P\left(c \qquad {}_j\middle|x\right)$ is the probability of unknown value of $class_j$ using the predictors in dataset x

The $P\left(x\middle|class_j\right)$ is the probability of predictors, given the $class_j$. That is consisted of training set

$P(c \qquad {}_j)$ is the probability of known value of $class_j$.

$P(x)$ is the evidence

The Naïve-Bayes function is placed in “e1071” package in R. The following algorithm shows this procedure.

```
Input: X, the set of training objects which is all predictors having value, the test objects including as z, which is partial attributes having value and L, the set of predictors having missing value that is used as class
Output: the predicted value of each class $l_i$
For each $l_i \in L$ do
Model =Naïve Bayes (X)
$l_i$=Naïve Bayes (Model, z)
```

4.6 Implementation

After the data is processed, it can be used for further analysis. To this aim, different steps of CBR have been implemented to retrieve the similar groups of patients and to adapt the best treatment for the new patient.

4.6.1 Case Retrieval

As mentioned before, the case Retrieval is the first step of CBR and it is used to find the most similar groups of patients. To this aim, the K-means clustering algorithm has been applied and the data is categorized into 20 different number of clusters.

There are different approached to find the optimum number of clusters. Some of them has been applied in this study which are including as:

- Partitioning Around Medoids (PAM): in Partitional Medoid Algorithm, the whole dataset is broken into groups and the distance between points labeled to be in a cluster and a point designated as the center of that cluster are defined to be minimized. The result is shown in Figure 4.2.

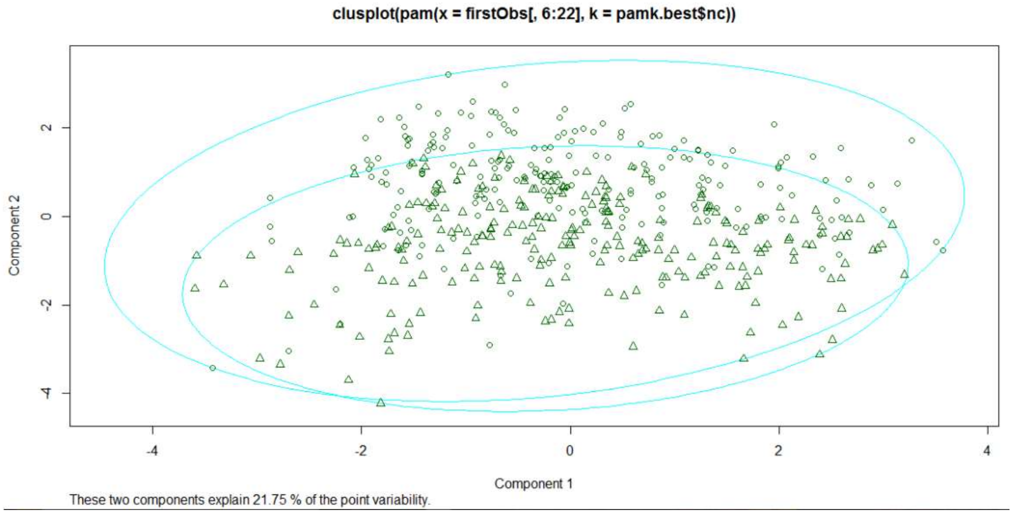


Figure 4. 2: PAM Clustering for Optimum Number of Clusters

Number of clusters:2

- Hierarchical Clustering: in this method, the hierarchy cluster model is built and based on two available approach “top down” and “bottom up”, the final number of clusters would be identified. The following result in Figure 4.3 defines the optimum number of clusters in this approach.

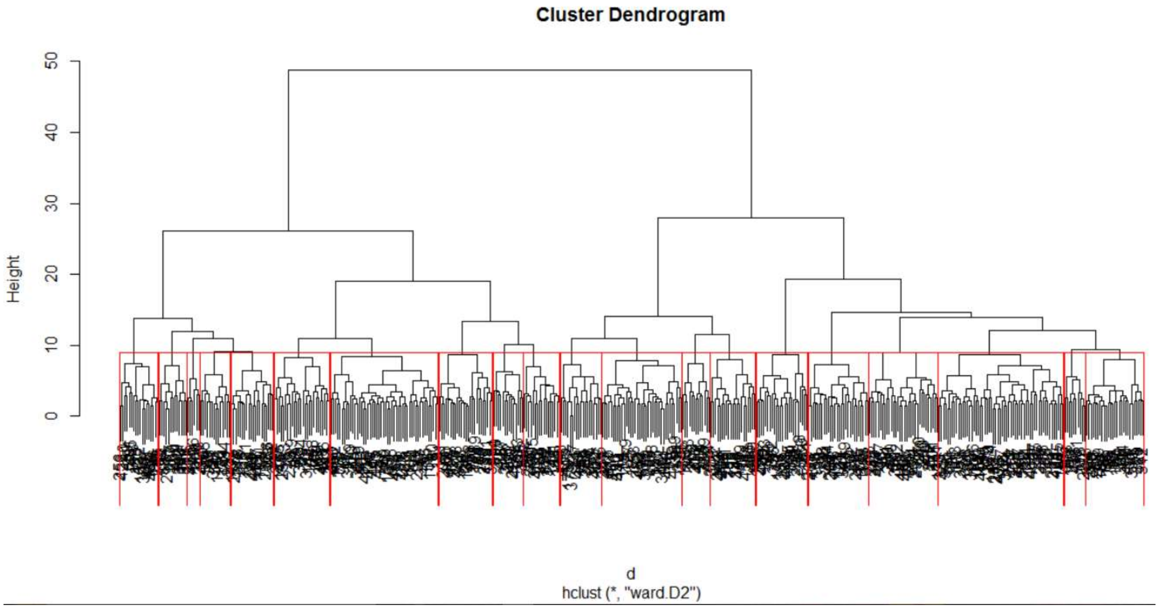


Figure 4. 3: Hierarchy Clustering for Optimum Number of Clusters

Optimum Number of clusters: 3

- Model Based Clustering: In this method, different models of data are generated and the maximum likelihood and Bayes criteria is estimated to define the most likely model and number of clusters. After applying this model, the optimum number of clusters is defined through Figure 4.4.

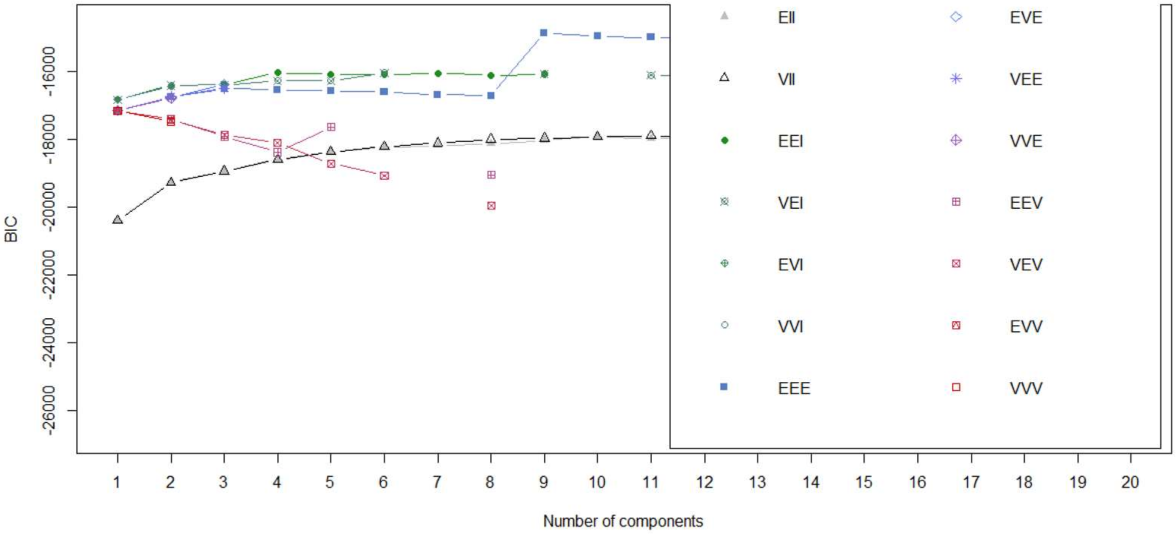


Figure 4. 4: Model Based Clustering for Optimum Number of Clusters

Optimum Number of clusters: 20

After comparing the different clustering analysis results, and “Within SSQ” and “Between SSQ” metrics, the result with the lowest “Within SSQ” and highest “Between SSQ” is selected for optimum number of clusters. Since the “Within SSQ” is showing the sum of squares between data points within the cluster, while “Between SSQ” shows the sum of squares of data points between the different clusters. In this case, the Model based clustering method has the lowest rate of error which is selected as the best result. This result is also showed in Table 4.10 and Figure 4.5.

Name of Approach	Number of clusters	Within SSQ Cluster	Between SSQ
PAM	2	6.22	20.7%
Model Based Cluster	20	0.622	60%
Hierarchical Cluster	3	4.14	34%

Table 4. 10: Comparing Within SSQ Cluster Error of Different Approaches

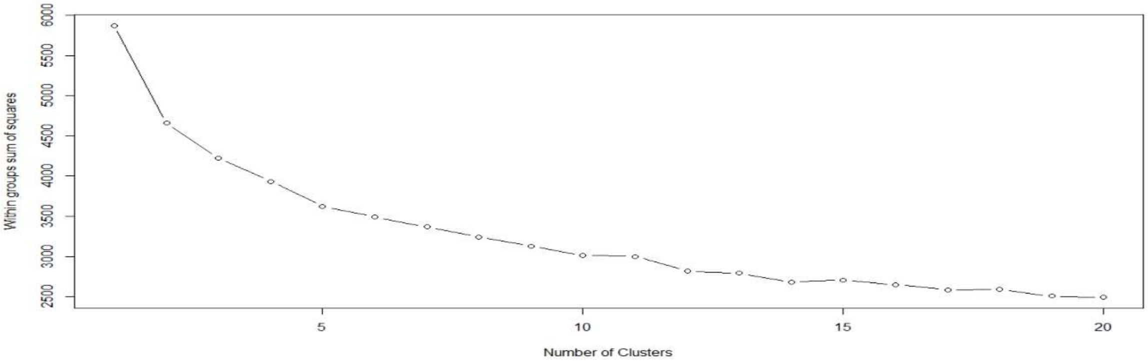


Figure 4. 5: Number of Clusters vs Within SSQ

After selecting the optimum number of clusters, the K-means clustering algorithms is used to cluster the First visit records of patients. Based on the obtained result, the similar patients are grouped in to 20 clusters that their subsequent records are used for further process.

4.6.2 Case Reuse

In this step, the subsequent records of similar patients in each cluster is extracted from the Subsequent dataset. Tracking the SBP level of each individual patient in a cluster, shows that the medication results were not successful all the time. In some cases, the increase of blood pressure after the visit is observed. Therefore, in order to have better results of treatment for new patient with similar condition, those unsuccessful records should be omitted and the treatment should be revised.

4-6-3 Case Revise

In order to revise the treatment of each patient, the successful result of each treatment should be defined. Figure 4.6 shows the various trends of blood pressure for 6 different patients. During different number of visits, the blood pressure doesn't have the descending slope all the time. So, in order to find the successful visits, the records that could reduce the blood pressure for the next visit or stabilize the blood pressure at the lowest level are selected as successful treatment.

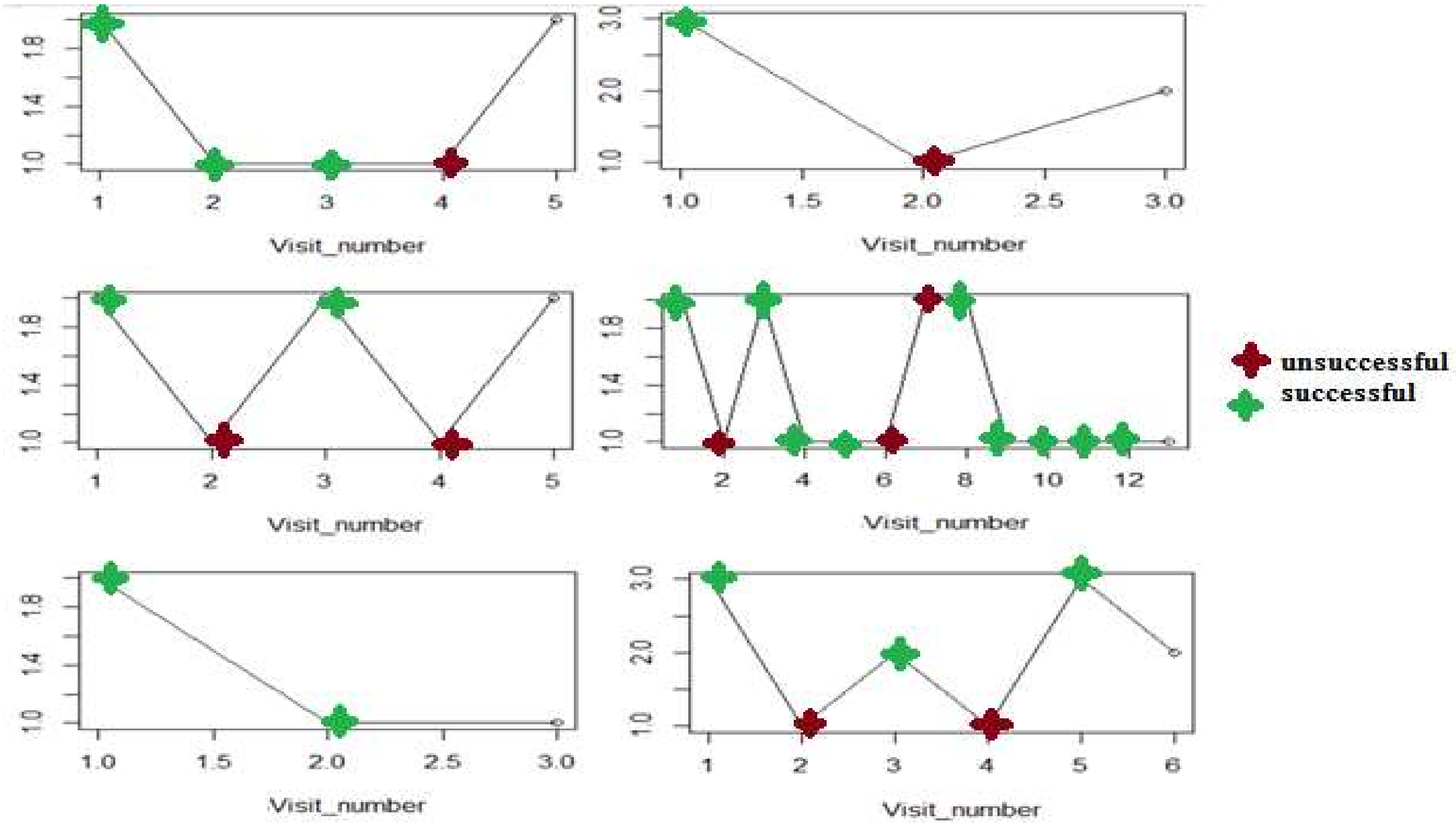
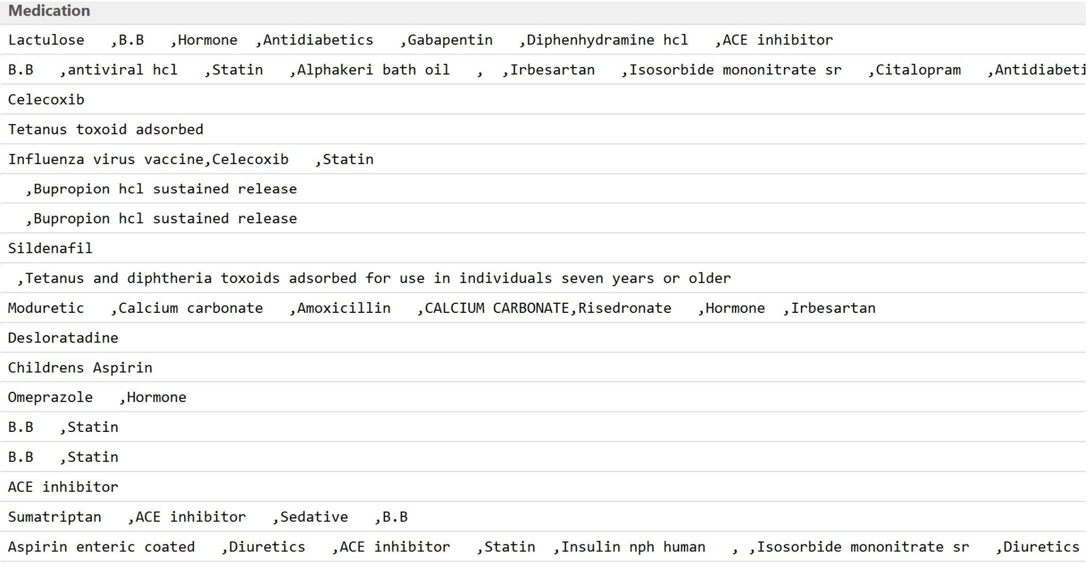


Figure 4. 6: Successful and Unsuccessful Visit

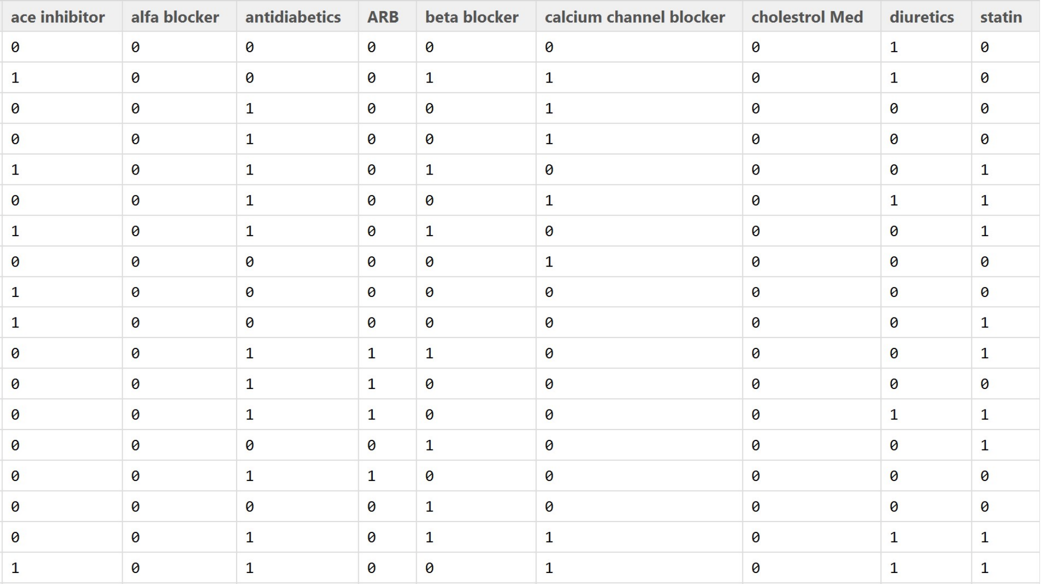
After selecting the successful visits for each individual group in cluster, the multi-label classification algorithm is used to classify the treatment of each group. The result will show the predicted probability of each drug that can be applied for the new patient.

Tables 4.11 and 4.12, show the sample of medication list before and after cleaning respectively.



Medication
Lactulose ,B.B ,Hormone ,Antidiabetics ,Gabapentin ,Diphenhydramine hcl ,ACE inhibitor
B.B ,antiviral hcl ,Statin ,Alphakeri bath oil , ,Irbesartan ,Isosorbide mononitrate sr ,Citalopram ,Antidiabeti
Celecoxib
Tetanus toxoid adsorbed
Influenza virus vaccine,Celecoxib ,Statin
,Bupropion hcl sustained release
,Bupropion hcl sustained release
Sildenafil
,Tetanus and diphtheria toxoids adsorbed for use in individuals seven years or older
Moduretic ,Calcium carbonate ,Amoxicillin ,CALCIUM CARBONATE,Risedronate ,Hormone ,Irbesartan
Desloratadine
Childrens Aspirin
Omeprazole ,Hormone
B.B ,Statin
B.B ,Statin
ACE inhibitor
Sumatriptan ,ACE inhibitor ,Sedative ,B.B
Aspirin enteric coated ,Diuretics ,ACE inhibitor ,Statin ,Insulin nph human , ,Isosorbide mononitrate sr ,Diuretics

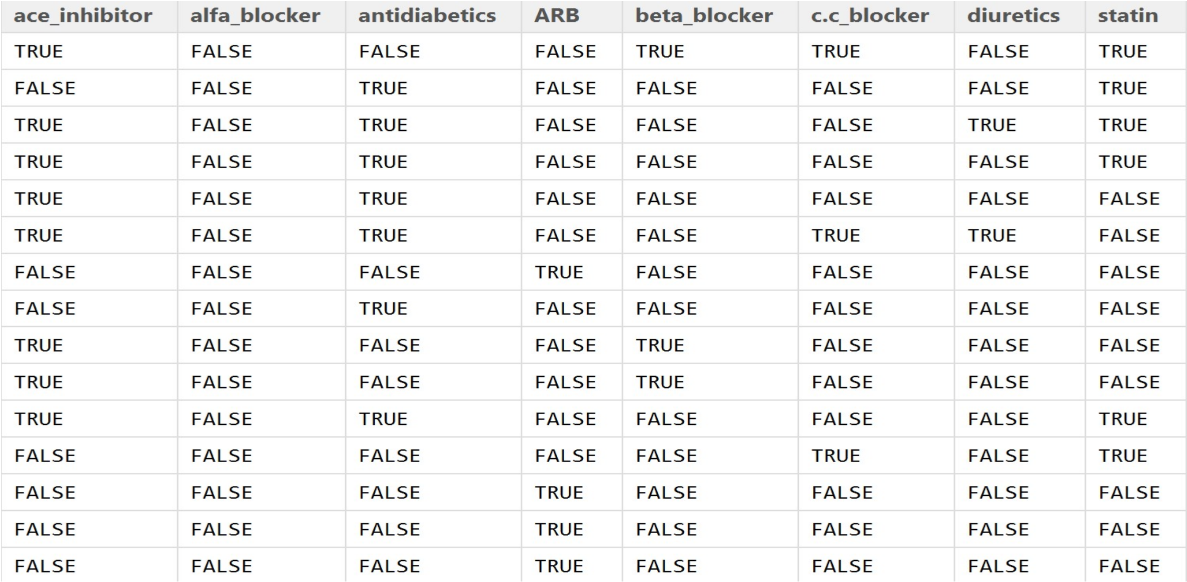
Table 4. 11: Medication Table Before Cleaning



ace inhibitor	alfa blocker	antidiabetics	ARB	beta blocker	calcium channel blocker	cholestrol Med	diuretics	statin
0	0	0	0	0	0	0	1	0
1	0	0	0	1	1	0	1	0
0	0	1	0	0	1	0	0	0
0	0	1	0	0	1	0	0	0
1	0	1	0	1	0	0	0	1
0	0	1	0	0	1	0	1	1
1	0	1	0	1	0	0	0	1
0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1
0	0	1	1	1	0	0	0	1
0	0	1	1	0	0	0	0	0
0	0	1	1	0	0	0	1	1
0	0	0	0	1	0	0	0	1
0	0	1	1	0	0	0	0	0
0	0	0	0	1	0	0	0	0
0	0	1	0	1	1	0	1	1
1	0	1	0	0	1	0	1	1

Table 4. 12: Medication Table After Cleaning

After compiling the successful visits of all patients into a table, the library “mlr” in R package classifies the data. First, the medication list is converted from binary to logical value. The logical list of those medication is shown in Table 4.13. Second, the function “MultiLabelTask” is used to find the classification task of the data. The Task model generates some information for multi-label classification learner such the type of the classification, the number of records, number of labels and the frequency of each label.



ace_inhibitor	alfa_blocker	antidiabetics	ARB	beta_blocker	c.c_blocker	diuretics	statin
TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE

Table 4. 13: Logical List of Medication

Input: The training set of successful records of different patients in a cluster with treatment information and Testing set of patients' records

Output: The predicted probability of treatment for testing set of records.

```
Task= makeMultilabelTask (data)
Train=train (Learner, Task, training set)
Predict= predict (Train, Task, testing set)
```

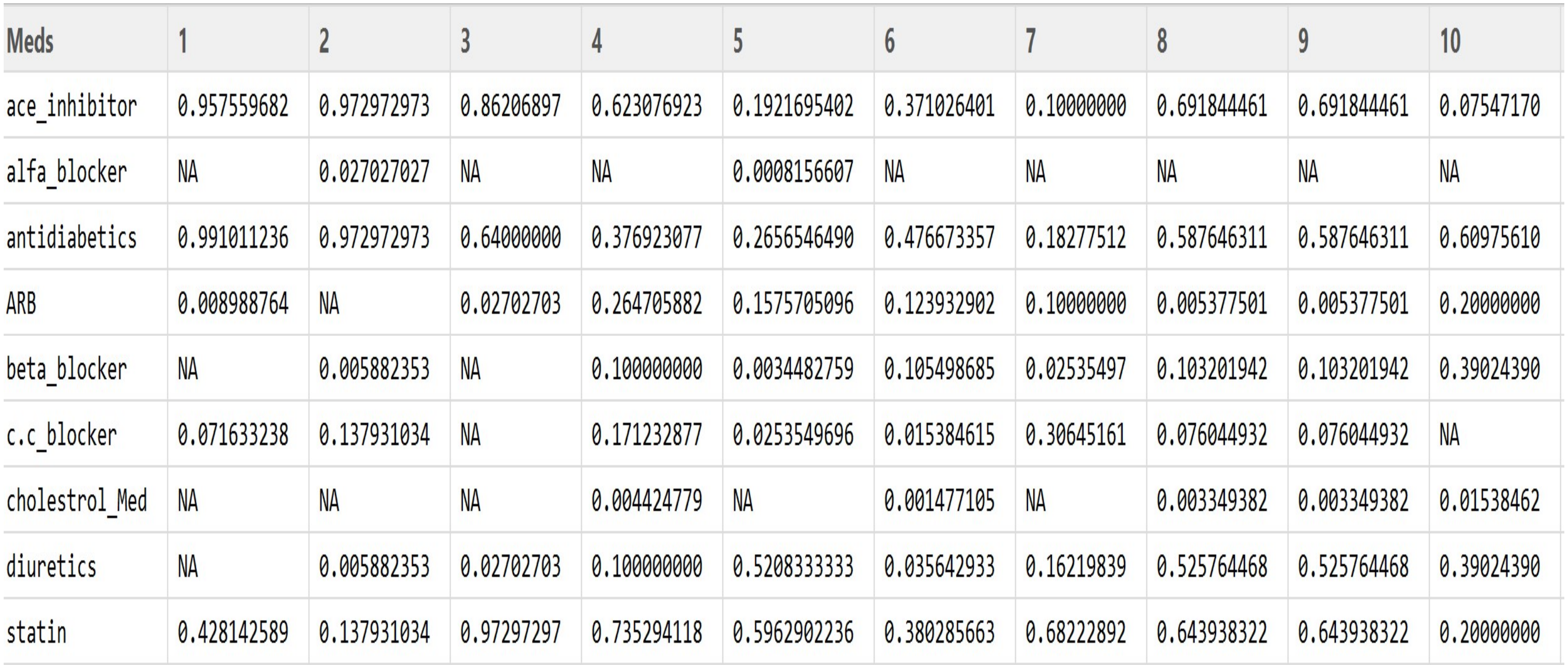
Besides the task generation, the multi-label classification learner should also be defined for further processing. The “makeLearner” function in “mlr” library is used to apply different types of learners' functions like "classif.lda", “classif.rpart” and “"multilabel.rFerns". Each learner has different model for multi-label classification:

- Classif.Ida function estimates the correlation between each attribute with respect to individual class.
- Classif.rpart uses decision tree algorithm to classify the multi-label problem
- rFerns or Random ferns is type of constraint classification decision tree. In this method the splitting criteria is done entirely randomly.

After defining the learner and applying the classification algorithm, the results should be combined to predict the probability of each label. The “makeMultilabelBinaryRelevanceWrapper” function is used to generate wrapping model of binary relevance classification results.

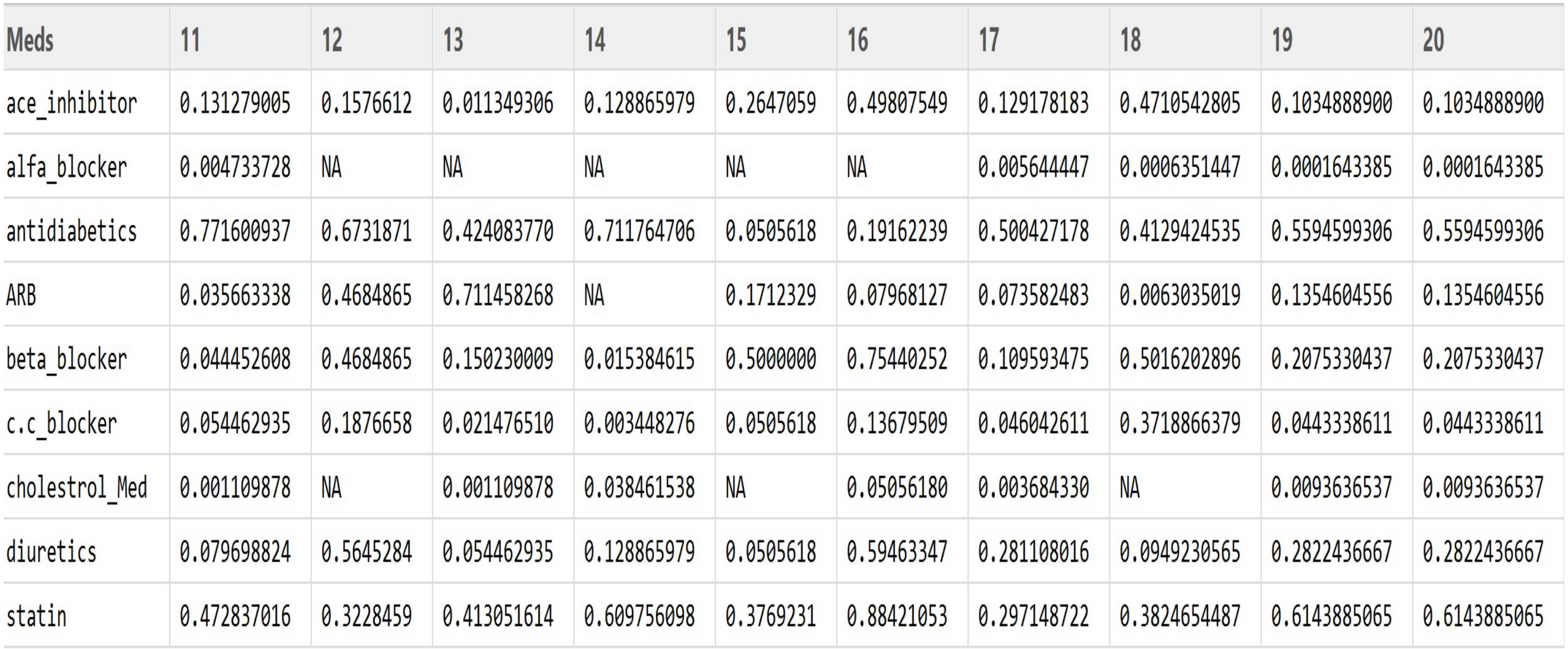
The next step after learning the multi-label classification is to train the dataset. To this aim, the successful visit dataset is exploited as training and testing records to build a prediction model. The “train” function in R is used to achieve this process.

Then, the “predict” function in R is employed to obtain the predicted medication results for the selected test cases of each cluster. Table 4.14 shows the different medication treatment results for each cluster.



Meds	1	2	3	4	5	6	7	8	9	10
ace_inhibitor	0.957559682	0.972972973	0.86206897	0.623076923	0.1921695402	0.371026401	0.10000000	0.691844461	0.691844461	0.07547170
alfa_blocker	NA	0.027027027	NA	NA	0.0008156607	NA	NA	NA	NA	NA
antidiabetics	0.991011236	0.972972973	0.64000000	0.376923077	0.2656546490	0.476673357	0.18277512	0.587646311	0.587646311	0.60975610
ARB	0.008988764	NA	0.02702703	0.264705882	0.1575705096	0.123932902	0.10000000	0.005377501	0.005377501	0.20000000
beta_blocker	NA	0.005882353	NA	0.100000000	0.0034482759	0.105498685	0.02535497	0.103201942	0.103201942	0.39024390
c.c_blocker	0.071633238	0.137931034	NA	0.171232877	0.0253549696	0.015384615	0.30645161	0.076044932	0.076044932	NA
cholestrol_Med	NA	NA	NA	0.004424779	NA	0.001477105	NA	0.003349382	0.003349382	0.01538462
diuretics	NA	0.005882353	0.02702703	0.100000000	0.5208333333	0.035642933	0.16219839	0.525764468	0.525764468	0.39024390
statin	0.428142589	0.137931034	0.97297297	0.735294118	0.5962902236	0.380285663	0.68222892	0.643938322	0.643938322	0.20000000

Table 4. 14: Classified Treatment for Each Cluster (1)



Meds	11	12	13	14	15	16	17	18	19	20
ace_inhibitor	0.131279005	0.1576612	0.011349306	0.128865979	0.2647059	0.49807549	0.129178183	0.4710542805	0.1034888900	0.1034888900
alfa_blocker	0.004733728	NA	NA	NA	NA	NA	0.005644447	0.0006351447	0.0001643385	0.0001643385
antidiabetics	0.771600937	0.6731871	0.424083770	0.711764706	0.0505618	0.19162239	0.500427178	0.4129424535	0.5594599306	0.5594599306
ARB	0.035663338	0.4684865	0.711458268	NA	0.1712329	0.07968127	0.073582483	0.0063035019	0.1354604556	0.1354604556
beta_blocker	0.044452608	0.4684865	0.150230009	0.015384615	0.5000000	0.75440252	0.109593475	0.5016202896	0.2075330437	0.2075330437
c.c_blocker	0.054462935	0.1876658	0.021476510	0.003448276	0.0505618	0.13679509	0.046042611	0.3718866379	0.0443338611	0.0443338611
cholestrol_Med	0.001109878	NA	0.001109878	0.038461538	NA	0.05056180	0.003684330	NA	0.0093636537	0.0093636537
diuretics	0.079698824	0.5645284	0.054462935	0.128865979	0.0505618	0.59463347	0.281108016	0.0949230565	0.2822436667	0.2822436667
statin	0.472837016	0.3228459	0.413051614	0.609756098	0.3769231	0.88421053	0.297148722	0.3824654487	0.6143885065	0.6143885065

Table 4. 15:Classified Treatment for Each Cluster (2)

Besides knowing the probability of predicted medications of each cluster, other information is including as:

The impact of using different antihypertensive medications for dropping the blood pressure to lower levels or stabilizing the blood pressure at level 1 according to these formulas:

$$Drop\ \% = \frac{freuency\ of\ medication\ x\ at\ SBP>1}{total\ number\ of\ records} \quad (4\text{-}3)$$

$$Stable\ \ \% = \frac{freuency\ of\ medication\ x\ at\ SBP=1}{total\ number\ of\ records} \quad (4\text{-}4)$$

After defining the successful visits, the frequency of different treatments with respect to the blood pressure level indicates the impact of each medication. Figure 4-7 shows the effect of different hypertension treatments for the sample cluster 1.

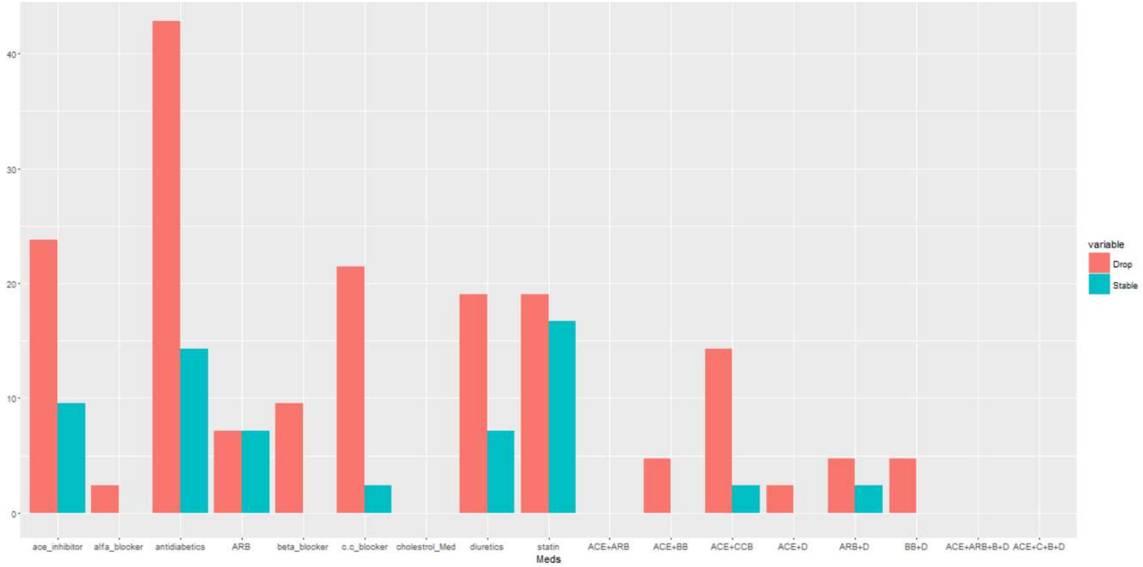


Figure 4. 7: The Impact of Different Hypertension Treatments

In addition to successful visits, the successful treatment is also used to define the following metrics:

- The impact of medication in blood pressure control given as

$$\boldsymbol{Successful\ Treatment\ Ratio} = \frac{\boldsymbol{\#\ of\ patients\ reach\ lowest\ BP\ for\ the\ last\ 2\ \ visits}}{\boldsymbol{Total\ number\ of\ \ patients\ in\ each\ cluster}} \quad (4\text{-}5)$$

- The shortest treatment path and average time of the treatment process given as

$$\boldsymbol{Shortest\ Treatment\ Path} = \boldsymbol{min}\{\ \boldsymbol{duration\ of\ successful\ treatments}\} \ (4\text{-}6)$$

Each similar patient, has different number of visits for their subsequent treatments. After selecting the patients who were treated successfully, the average number of treatment or the average time of their treatment from starting to the end of the process would show the predicted treatment time for the new patient as well.

Chapter 5: Results and Discussion

5.1 Results

CBR classifies the patients into similar groups and revise the treatment of each cluster to propose the successful treatments. Having this information, provides the personalized treatment procedure for the new patient. The information of new patient is compared with other records in First visits dataset by using the similarity measurement.

There are many approaches for measuring the similarity. The general form of similarity measurement is to find the distance between two variables. There are different types of distance metric that are used for numerical variable such as Euclidean and Manhattan Distance. However, in this study, since some of the variables are categorical such as Race, Gender, Diagnosis type, Marital status, all other attributes are converted into categorical variable and the similarity measurement of categorical variables is applied. The easiest approach to compare two categorical variables is to assign value 1 when they match and 0 if they do not. Therefore, for multivariate categorical data, the sum of number of matches would indicate the similarity measurement. The disadvantage of this method is to ignore the probability of occurrence of each level. More information about the variables would be needed to have more accurate similarity measurement. Hence, for each attribute, the frequency of each level would be calculated to find the probability occurrence of each level (Boriah, et al. 2016).

Generally, based on Boriah (2016) study, different factors infer the similarity measure of categorical variables which are included as N (number of records) d (number of attributes) k (number of levels of each attribute) and f (the frequency of each level).

In this study, the Occurrence frequency (OF) method has been applied to find the similarity measurement of the data. The inconformity between less frequent values refers to the less similarity while the inconformity between frequent values refers to high similarity.

Since the frequency levels of each attribute weren't the same, this similarity approach would have better result. In general, the similarity function is defined through the following equation:

$$sim = \frac{1}{1+dist} \tag{5-1}$$

Where dist is referring to the distance (difference) between two cases. The total similarity between two records is calculated by using the following equation:

$$S(X,Y) = \sum_{k=1}^{d} w_k S_k(X_k, Y_k) \tag{5-2}$$

Where, S_k (X_k,Y_k) represents the similarity between record X and record Y for each attribute and w_k is the weight of each attribute and d is the total number of attributes.

Based on OF method, the following equation is used to define the similarity between each attribute:

$$S_k(X_k, Y_k) = \begin{cases} 1 & \text{if } X_k = Y_k \\ \frac{1}{1+\log \frac{N}{f_k(X_k)} \times \log \frac{N}{f_k(Y_k)}} & \text{otherwise} \end{cases}$$

(5-3)

When the attributes are having the same value, then the similarity is equal to 1. In case of inequality of attributes, the log formula has been applied to find the similarity measurement between cases. The N and f_k (x_k) would be the total number of records and the frequency of value x for attribute k respectively.

Besides the frequency value, the weight of each attribute should be defined. The logistic regression analysis has been used to defined the weight of each features, based on the binary drug labels.

In this study, nine different medications have been used as class labels. Therefore, in order to find the final weights of the attributes, the logistic regression has been applied for each individual class label and then the average weights of each feature would be represent the final weights. This procedure is shown in Figure 5.1.

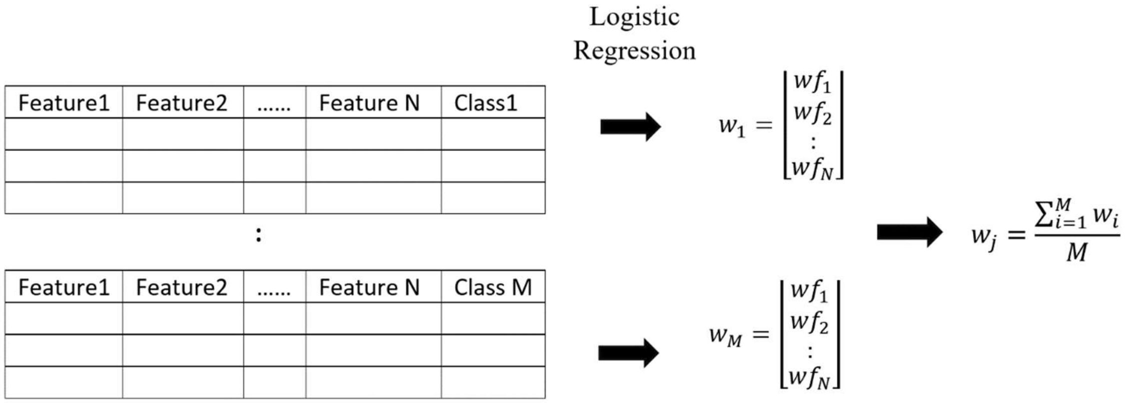

Logistic Regression
Feature1
Feature2
......
Feature N
Class1
:
Class M
$w_1 = \begin{bmatrix} wf_1 \\ wf_2 \\ \vdots \\ wf_N \end{bmatrix}$
$w_M = \begin{bmatrix} wf_1 \\ wf_2 \\ \vdots \\ wf_N \end{bmatrix}$
$w_j = \frac{\sum_{i=1}^{M} w_i}{M}$

Figure 5. 1: Finding the Feature Weight Using Logistic Regression

The library “MASS” in R is used to apply the “glm” function to find the weight of the features by using logistic regression algorithm.

```
Input: The First visits dataset, including X, the set of features and L, the set of medication labels
Output: The weight of each feature with respect to the L
For each $l_j$ є L
$w_j$=logistic regression (X,$l_j$)
W= average ($w_{j)}$
```

So, when the new patient comes, the similar matches are selected by calculating the similarity between all the records. Then, the top 3 closest matches are used for further process.

After the similar matches are selected, based on their assigned cluster, the treatment result is applied to the new patient. In this situation, different results might be achieved.

- Similar results belong to the same cluster
- Similar results belong to different cluster

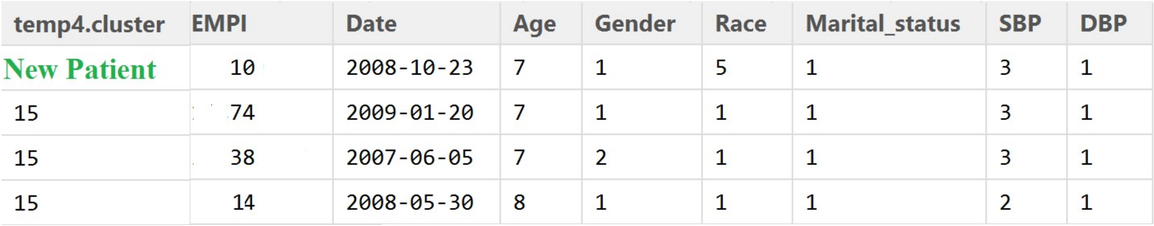
In first situation, when all cases are belonged to the same cluster, the treatment of that cluster is directly applied to the new patient.

However, for the second situation, due to overlap of multi clusters, the treatment plan is not straight forward and the subscription of predicted treatment is applied as the treatment result.

In order to define the best treatment among all different drug types, 3 of the most frequent medications are selected as assigned treatment for the cluster.

The following examples are showing the different results of similarity measurement.

Test Case 1:

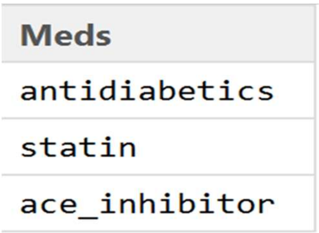


temp4.cluster	EMPI	Date	Age	Gender	Race	Marital_status	SBP	DBP
New Patient	10	2008-10-23	7	1	5	1	3	1
15	74	2009-01-20	7	1	1	1	3	1
15	38	2007-06-05	7	2	1	1	3	1
15	14	2008-05-30	8	1	1	1	2	1

Table 5. 1: The Similarity Result of Case1

As it can be seen from the Table 5.1, all similar cases belong to the same cluster. So, the treatment result would be the revised treatment of cluster 15. Table 5.2 shows the treatment result of that cluster.

Based on the result of treatment for cluster 15, “ACE inhibitor”, “Antidiabetics” and “Statin” are the most frequent medications for that cluster that can be applied for the new patient.

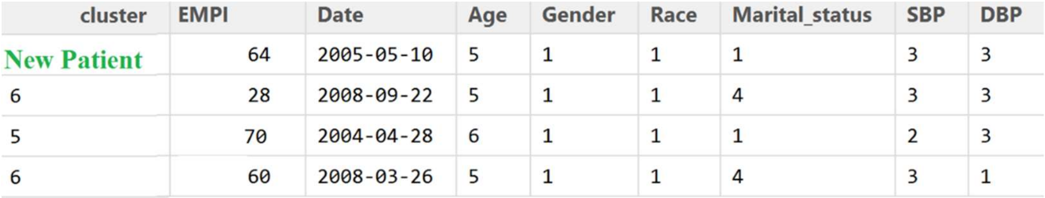


Meds
antidiabetics
statin
ace_inhibitor

Table 5. 2: The Treatment Result of Case1

Test Case 2:

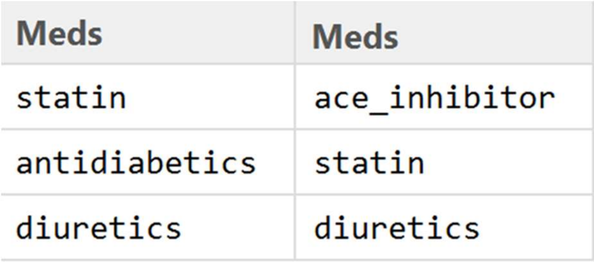
In this case, the similar records belong to the 2 different clusters 18 and 7. The similarity result and the associated drug treatment is shown in Tables 5.3 and 5.4 respectively. So, the subscription of these treatments are selected for the new patient treatment.



cluster	EMPI	Date	Age	Gender	Race	Marital_status	SBP	DBP
New Patient	64	2005-05-10	5	1	1	1	3	3
6	28	2008-09-22	5	1	1	4	3	3
5	70	2004-04-28	6	1	1	1	2	3
6	60	2008-03-26	5	1	1	4	3	1

Table 5. 3:The Similarity Result of Case2

The treatment result of cluster 5 indicates that, “ACE inhibitor”, “Statin” and “diuretics” are the most frequent medications while in cluster 6, “Statin”, “antidiabetics” and “diuretics” are known as the selected treatment. Therefore, in this case, the “diuretics” and “Statin” are the selected medications.

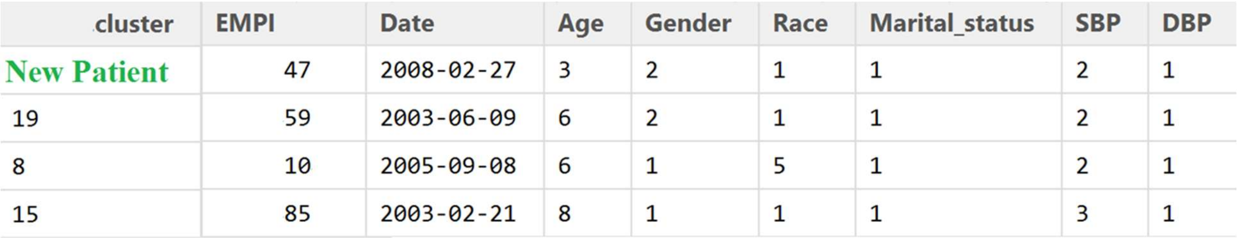


Meds	Meds
statin	ace_inhibitor
antidiabetics	statin
diuretics	diuretics

Table 5. 4:The Treatment Result of Case2

Test Case 3:

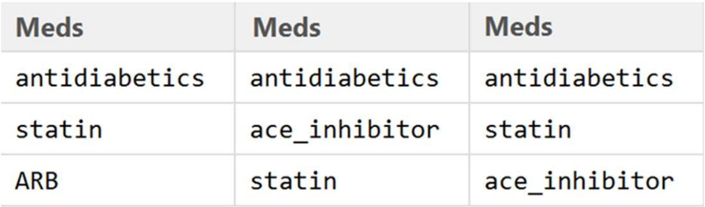
In this case, the similar candidates belong to 3 different clusters. So in order to find the final treatment, the assigned medication of clusters should be compared. Tables 5.5 and 5.6 show those results.



cluster	EMPI	Date	Age	Gender	Race	Marital_status	SBP	DBP
New Patient	47	2008-02-27	3	2	1	1	2	1
19	59	2003-06-09	6	2	1	1	2	1
8	10	2005-09-08	6	1	5	1	2	1
15	85	2003-02-21	8	1	1	1	3	1

Table 5. 5: The Similarity Result of Case3

Based on the treatment result of different clusters, “Antidiabetics” and “Statin” and “ARB” from cluster 19, “antidiabetics”, “ace-inhibitor” and “statin” from cluster 8 and “ACE inhibitor”, “Antidiabetics” and “Statin” from cluster 15 are selected as best treatment. However, based on subscription of different clusters, the “Statin”, “antidiabetics” and “ace-inhibitors” are known as the best treatment for new patient.



Meds	Meds	Meds
antidiabetics	antidiabetics	antidiabetics
statin	ace_inhibitor	statin
ARB	statin	ace_inhibitor

Table 5. 6: The Treatment Result of Case3

5.2 Validation

In order to find the accuracy and reliability of this technique, the results should be validated. There are two available approaches in this regards. That are including as:

- The current approach: Since in this study, the new patient is a random selected records form the dataset, the medication result is available. Comparing the result of Temporal CBR with the existing treatment could validate this approach.

  To this aim, three different sets of data are selected as new case group. The proposed result of each case is compared with the actual probability value of prescribed medication. Then the final result of all test cases is defined as the accuracy of this algorithm.

  The used datasets are as:

  D1: Groups of patients with more than 5 visits (131 out of 475)

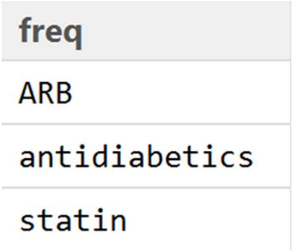
  D2: Groups of patients with more than 10 visits (45 out of 475)

  D3: Groups of patients with more than 2 visits AND Successful treatment (66 out of 475)

  P1: probability of new case records having similar treatment result with proposed algorithm with accuracy > 0.5

  P2: probability of new case records having similar treatment result with proposed algorithm with accuracy >0.6

  As an example, considering the test case 3. This case is a patient who belongs to D3 group of data. The proposed medical solution is shown in Table 5.7. The probability of actual medication after selecting the top 3 frequent drugs is shown in Table 5.8.



freq
ARB
antidiabetics
statin

Table 5. 7: The Actual Treatment Result for Case 3

As can be seen from this table, the ARB, antidiabetics and Statin are the most frequent used medication for this patient. So by comparing those result, 7 out of 9 possible proposed drug is referring to the actual treatment. Therefore, the accuracy of proposed treatment for this test case is equal to 77.77%. The result of reiterating this procedure for all the new cases in different dataset is shown in Table 5.8.

	P1	P2
D1	80%	59%
D2	86%	80%
D3	84%	60%

Table 5. 8: The validation Results of different datasets

The results of different datasets, show that for all group of cases, more than 80/5 of records have matching medication with their proposed result with accuracy higher than 0.5. However, the more records of patient available, the more convergence of proposed and actual treatment result would be achieved.

- Besides using current validation approach, there is another way to validate the result which can be done as a future works and that is using the confirmation the experts directly.

Chapter 6: Conclusion

6. Conclusion

In this study, the personalized treatment for hypertension is applied by using the temporal Case Based Reasoning (CBR). Screening the different medical test records of hypertensive patients and reviewing medical surveys has led to select different patients' parameters as significant risk factors to tailor the medication. Those parameters include as demographic information, physical examination and laboratorial test results. Using this information to build a temporal CBR model, has led to cluster the similar patients in case base and group the treatment approach for each cluster. So, when the new patient comes, the similarity between new case and other records in case base is compared and the top 3 candidates are selected as the closest match. The results show that those candidates might either belong to the same cluster or different cluster. When the candidates are having the same cluster, the treatment of that cluster is directly applied to the new patient. However, in case of having candidates with different cluster, the subscription of clusters is used as a final treatment for the new patient. Applying successful treatment results of previous patients to build a treatment model and selecting the most similar candidate as a reference of treatment, has enabled this approach as an efficient tool to provide the personalized treatment for the new patient.

Moreover, this method would provide physicians with some other useful metrics such as successful treatment ratio and the shortest treatment path for each group of medication that provides a useful guidance to the proposed treatment approach.

7. Reference

Aamodt, A. and Plaza, E. (1994). “Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches” 7:1: 39–59.

Cordier, A. and Fuchs, B. (2007). “Failure Analysis for Domain Knowledge Acquisition in a Knowledge-Intensive CBR System.,” 463–77. doi:10.1007/978-3-540-74141-1_32.

Mayo Clinic staff. (2016). “Angiotensin II Receptor Blockers.” *RxList*

Chebel-Morello, B. and Haouchine, M. K. (2013). “Reutilization of Diagnostic Cases by Adaptation of Knowledge Models.” *Engineering Applications of Artificial Intelligence* 26 (10): 2559–73. doi: 10.1016/j.engappai.2013.05.001.

Byrd, J.B. (2016). “Personalized Medicine and Treatment Approaches in Hypertension: Current Perspectives.” *Integrated Blood Pressure Control* 9 (April): 59–67. doi:10.2147/IBPC.S74320.

Castro, J. L., Navarro, M., Sánchez, J. M. and Zurita, J. M. (2009). “Loss and Gain Functions for CBR Retrieval.” *Information Sciences*, Including Special Issue on Chance Discovery of Significant Events for Decision, 179 (11): 1738–50. doi: 10.1016/j.ins.2009.01.017.

Brauser, D. (2015). “CDC: Hypertension-Related Mortality Has Climbed Since 2000.”

Chanumin, M. (2008). “Study on Case Retrieving in Case-Based Reasoning Based on Grey Incidence Theory and Its Application in Bank Regulation.” *FUZZ-IEEE 2008*, no. Proceedings of Fuzzy Systems: 1530–33.

Umamaheswari, D. and Shyamala, N. (2015). "Data Mining Techniques to Fill the Missing Data and Detecting Patterns.” IJSTE - International Journal of Science Technology & Engineering 2 (1), ISSN (online): 2349-784X

Leake, D. B. (1996). “CBR in Context: The Present and Future.”

Patterson, D. W., Rooney, N. (2002). “A Regression Based Adaptation Strategy for Case-Based Reasoning.,” 87–92.

WebMD. (2016). “Diabetes and High Blood Pressure.” http://www.webmd.com/hypertension-high-blood-pressure/guide/high-blood-pressure.

Dombawalage, F., Irosh, A. and Henskens, F. A.(2015). “A Modified Case-Based Reasoning Approach for Triaging Psychiatric Patients Using a Similarity Measure Derived from Orthogonal Vector Projection.” In *Artificial Life and Computational Intelligence*, edited by Stephan K. Chalup, Alan D. Blair, and Marcus Randall, 360–72. Lecture Notes in Computer Science 8955. Springer International Publishing. http://link.springer.com/chapter/10.1007/978-3-319-14803-8_28.

Alan, S., Bauman, M.A., Sallyann, M., King, C., Fonarow, G. C., Lawrence, W., Williams, K. A., and Sanchez, E. (2013). “An Effective Approach to High Blood Pressure Control. A Science Advisory from the American Heart Association, the American College of Cardiology, and the Centers for Disease Control and Prevention.” *Hypertension*, HYP.0000000000000003. doi:10.1161/HYP.0000000000000003.

Charis, G. (2016). “High Blood Pressure & Triglycerides.” *LIVESTRONG.COM*. Accessed June 13. http://www.livestrong.com/article/431399-high-blood-pressure-triglycerides/.

Lichcuan,G., Guo, Q., Cao, M. and Zhang, D. (2013). “Case Retrieval Algorithm Based on Structure Matrix.” *Journal of Multimedia* 8 (6): 712–19. doi:10.4304/jmm.8.6.712-719.

Hamburg, M. A., and Francis S. C. (2010). “The Path to Personalized Medicine.” *New England Journal of Medicine* 363 (4): 301–4. doi:10.1056/NEJMp1006304.

Sánchez M. M., Cortés U., Martínez M., Comas. J., and Rodríguez. R, I. (2005). "An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning." *Case-Based Reasoning Research and Development*, Volume 3620: 465-476

Hassin, M. H. B. M., Norwawi, N. B. M. and Aziz, A. B. A. (2006). "Temporal Case-Based Reasoning for Reservoir Spillway Gate Operation Recommendation." In *2006 International Conference on Computing Informatics*, 1–4. doi:10.1109/ICOCI.2006.5276517.

Hensing, T., Apoorva C., Rishi B., and Ravi S. (2014). "A Personalized Treatment for Lung Cancer: Molecular Pathways, Targeted Therapies, and Genomic Characterization." *Advances in Experimental Medicine and Biology* 799: 85–117. doi:10.1007/978-1-4614-8778-4_5.

Hsu, C.C., and Ho, C.S. (2004). "A New Hybrid Case-Based Architecture for Medical Diagnosis." *Information Sciences* 166 (1–4): 231–47. doi: 10.1016/j.ins.2003.11.009.

Hu, J. J.Q., and Yinghong, P. (2015). "New CBR Adaptation Method Combining with Problem–solution Relational Analysis for Mechanical Design." *Computers in Industry* 66 (January): 41–51. doi: 10.1016/j.compind.2014.08.004.

Huang, M.J., Chen, M.Y. and Lee, S.C. (2007). "Integrating Data Mining with Case-Based Reasoning for Chronic Diseases Prognosis and Diagnosis." *Expert Systems with Applications* 32 (3): 856–67. doi: 10.1016/j.eswa.2006.01.038.

Hugh, R. O., Bridge, D. G. (2005). "A Case Base Similarity Framework" Volume 1168 (June): pp 309-323.

Hui, D. (2009). "An Improving Method of CBR Retrieval Based on Self-Organizing Map." doi:10.1109/ICICISYS.2009.5357621.

Kolodner, J. (2014). "*Case Based Reasoning",* IBPC-74320-Personalized-Medicine-and-Treatment-Approaches-in-Hypertension.

Jung, S., Taesoo L., and Dongsoo K. (2009). "Integrating Radial Basis Function Networks with Case-Based Reasoning for Product Design." *Expert Systems with Applications* 36 (3, Part 1): 5695–5701. doi: 10.1016/j.eswa.2008.06.099.

Hanney, K., Keane, M.K. (1996). "Learning Adaptation Rules from a Case-Base.," 179–92. doi:10.1007/BFb0020610.

Kapil, K. and Sharma, D.P. (2015). "Hybrid Reasoning Model for Strengthening the Problem Solving Capability of Expert Systems." *International Journal of Advanced Computer Science and Applications(IJACSA)* 4 (10).

King, R. H., Stottler, J. A. (1989). "Rapid Retrieval Algorithms for Case-Based Reasoning."

Koton, P. (1988). "Reasoning about Evidence in Causal Explanations.," 256–63.

Laragh, J.H., Ulick, S., Januszewicz, V., Deming, Q.B., Kelly, W.G., Lieberman, S. (1960). "Aldosterone Secretion and Primary and Malignant Hypertension." 1091–1106.

Leake, D.B., Kinley, A. and Wilson, D. (1995). "Learning to Improve Case Adaptation by Introspective Reasoning and CBR." In *Case-Based Reasoning Research and Development*, edited by Manuela Veloso and Agnar Aamodt, 229–40.

Lewiston, N.Y. (1992). "Klassifikationsverfahren in Expertsystemen Fur Die Medizin." *Mellen University Press*.

Li, Y. F., Xie, M. and Goh, T. N. (2009). "A Study of Mutual Information Based Feature Selection for Case Based Reasoning in Software Cost Estimation." *Expert Systems with Applications* 36 (3, Part 2): 5921–31. doi: 10.1016/j.eswa.2008.07.062.

Martínez, M. C., Campos, M., Sainte-Maure, M., Comas, J. and Rodríguez-Roda, I. (2006). "Improving the Efficiency of Case-Based Reasoning to Deal with Activated Sludge Solids Separation Problems." *Environmental Technology* 27 (6): 585–96. doi:10.1080/09593332708618679.

Rudradeb, M., and Basak, J. (2005). "Methods of Case Adaptation: A Survey." *International Journal of Intelligent Systems* 20 (6): 627–45. doi:10.1002/int.20087.

Rezvan, M.T., Hamadani, A. Z., and Shalbafzadeh, A., Safari, B. (2014). "A Combined Data Mining Approach Using Rough Set Theory and Case Based Reasoning in Medical Datasets."

Montani, S., Bottrighi, A., Leonardi, G. and Portinale, L. (2009). "A CBR-Based, Closed-Loop Architecture for Temporal Abstractions Configuration." *Computational Intelligence* 25 (3): 235–49. doi:10.1111/j.1467-8640.2009.00340. x.

Peng, Y.T., Lin, C.Y., Sun, M.T. and Tsai, K. C. (2009). "Healthcare Audio Event Classification Using Hidden Markov Models and Hierarchical Hidden Markov Models." In , 1218–21. IEEE. doi:10.1109/ICME.2009.5202720.

Qi, J., Hu, J., and Peng, Y. (2012). "A New Adaptation Method Based on Adaptability under K-Nearest Neighbors for Case Adaptation in Case-Based Design." *Expert Systems with Applications* 39 (7): 6485–6502. doi: 10.1016/j.eswa.2011.12.055.

Qi, J., Hu, J., and Peng, Y. (2015). "Incorporating Adaptability-Related Knowledge into Support Vector Machine for Case-Based Design Adaptation." *Engineering Applications of Artificial Intelligence* 37 (January): 170–80. doi: 10.1016/j.engappai.2014.09.010.

Rainer, S. and Gierl, L. (2001). "Case-Based Reasoning for Antibiotics Therapy Advice: An Investigation of Retrieval Algorithms and Prototypes." *Artificial Intelligence in Medicine* 23 (2): 171–86. doi:10.1016/S0933-3657(01)00083-5.

Winters-Miners, L.A. (2016). "Seven Ways Predictive Analytics Can Improve Healthcare."Elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare

Begum, S., Ahmed, M, U. (2009). "A Case-Based Decision Support System for Individual Stress Diagnosis Using Fuzzy Similarity Matching." *Computational Intelligence* 25: 180–95. doi:10.1111/j.1467-8640.2009.00337. x.

Krishnaswamy, S., Kang, Y.B. and Zaslavsky, A. (2014). "A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge."

Boriah, S., Kumar, V. (2016). "Similarity Measures for Categorical Data: A Comparative Evaluation." doi: http://dx.doi.org/10.1137/1.9781611972788.22

Dalal, S., Athavale, V. (2011). "Case Retrieval Optimization of Case-Based Reasoning through Knowledge-Intensive Similarity Measures."

Fanoiki, T. O., Drummond, I. (2010). "Case-Based Reasoning Retrieval and Reuse Using Case Resemblance Hypergraphs," 1–7. doi:10.1109/FUZZY.2010.5584854.

Turner, S.T., Gary, L. S., and Boerwinkle, E. (2007). "Personalized Medicine for High Blood Pressure." *Hypertension* 50 (1): 1–5. doi:10.1161/hypertensionaha.107.087049.

Wenkebach, U., Pollwein, B. (1992). "Visualization of Large Datasets in Intensive Care." *Proceedings / the ... Annual Symposium on Computer Application [Sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 18–22.

Weber, M. A., Schiffrin, E.L., William, B. W., Mann, S., Lindholm, L. H., Kenerson, J.G., Flack, J.M., et al. (2014). “Clinical Practice Guidelines for the Management of Hypertension in the Community: A Statement by the American Society of Hypertension and the International Society of Hypertension.” *The Journal of Clinical Hypertension* 16 (1): 14–26. doi:10.1111/jch.12237.

Petra, W., Deserno, T. M., Fischer, B., Günther, R.W.and Spreckelsen, C. (2011). “Towards Case-Based Medical Learning in Radiological Decision Making Using Content-Based Image Retrieval.” *BMC Medical Informatics and Decision Making* 11 (1): 1–16. doi:10.1186/1472-6947-11-68.

Montani, S., Bottrighi, A., Leonardi, G., Portinal, L. (2009). “A Cbr-Based, Closed-Loop Architecture for Temporal Abstractions Configuration.” Computational Intelligence 25(3): 235-249. doi: 10.1111/j.1467-8640.2009.00340.x.