

Evaluation of Features for Prediction of Hospitalization in Patients with End Stage Renal Disease

Elin Edvinsson, Anna Goos

28th May 2018



LUND UNIVERSITY

Master's Thesis in Biomedical Engineering
Faculty of Engineering, LTH
Department of Biomedical Engineering

Supervisor: Martin Stridh, LTH
Mattias Sellin, Lytics Health AB

Examiner: Frida Sandberg, LTH

Abstract

End Stage Renal Disease (ESRD) is a chronic kidney disease which results in lifelong care and regular dialysis treatments. It is both an expensive and extensive diagnosis that causes a lot of suffering for the patients. The workload on the healthcare system is increasing drastically, due to a population that lives longer with more chronic diseases. In this project, input data to an AI system was investigated and changed. The reviewed system predicts whether patients will be hospitalized in the near future or not. The input data consisted of features based on patients' medical histories and the purpose of this project was to manually create features with new or modified information. This was done in order to improve the system's performance with help of experiences from healthcare professionals. A test environment was used to evaluate the performance of the new feature sets. The newly created feature sets resulted in better predictions in some cases and worse in others compared to the original feature set. The results were measured with a Z-test with 95% confidence interval and the features were compared through a feature importance method. This project has shown that it is possible to improve AI predictions with the use of experiences from healthcare professionals during feature extraction. Since the improvements were just slightly better, the conclusion from this project was that larger structural changes are probably needed in order to convincingly increase the AI system's prediction performance.

Keywords: AI, ESRD, Features, Random forest

Acknowledgements

This thesis could not have been done only by us and luckily we had a lot of supporting people around us. To name and thank all of you would be impossible, but some have been especially important for us during this project. First of all, we want to thank all the people working at Lytics Health AB for taking care of us and making us feel welcome at the office, every day! Daniel and Mattias, your opinions and thoughts have been great, thanks for all your help. Thank you Axel, you are an excellent explainer and you helped us during the entire project with your computer skills. We also want to thank Jonas, Wendy, and Joann, whose expertise in this domain are priceless. Martin, our supervisor at the Department of Biomedical Engineering, you deserve a big thank for your important opinions during this work. We would probably not be able to complete this master's thesis without your invaluable help which helped us stay in the right direction during the entire project as well as your thoughts about the future as engineers. Thank you all!

Contents

1	Introduction	1
1.1	Abbreviations	3
2	Background	5
2.1	End Stage Renal Disease	5
2.2	Dialysis	7
2.2.1	Hemodialysis	7
2.2.2	Peritoneal Dialysis	7
2.3	Comorbidities	8
2.4	Medical Coding	9
2.5	ESRD Seamless Care Organization	10
2.6	Computer-based Decision Support System	10
2.6.1	Random Forest	11
2.7	Feature Extraction	12
2.7.1	Feature Construction	12
2.7.2	Feature Selection	13
2.8	ROC Curve	14
3	Method	17
3.1	Identification	18
3.1.1	The AI System	19
3.1.1.1	Input Data	20
3.1.1.2	Output Data	20
3.1.1.3	Missing Data	21
3.1.2	Test Environment	21
3.2	Best Practice	22
3.3	The Features	25
3.3.1	Feature Importance Method	26
3.3.2	Medication	27
3.3.2.1	Administered Medicines	28
3.3.2.2	Prescribed Medicines	29
3.3.2.3	Merged Medication Features	29
3.3.3	Dialysis	30
3.3.3.1	Blood Pressure	30
3.3.3.2	Blood Flow Rate	32
3.3.3.3	Post Weight Difference	34

3.3.4	Time Since Healthcare Event	35
3.3.4.1	Number of Admissions	35
3.3.4.2	Hospitalizations per Week	35
3.3.4.3	Top Three Time Since Healthcare Event	35
3.3.5	Laboratory Results and Diagnoses	36
3.3.5.1	Diagnoses	36
3.3.5.2	Age	37
3.3.6	Removing Features	38
3.4	Test Engineering	39
3.5	Validation	42
3.5.1	Prove Significance	42
3.5.2	Balanced Data Set	43
3.5.3	Influence the Results	44
4	Results _____	45
4.1	Identification	45
4.2	Best Practice	46
4.3	Feature Importance	48
4.4	Test Results	52
4.5	Validation	53
4.5.1	Prove Significance	53
4.5.2	Distribution of the Patients	54
4.5.3	Balanced Data Set	57
4.5.4	Influence the Results	57
5	Discussion _____	61
5.1	The New Features	61
5.2	Z-test	63
5.3	The Variation of the AUC Values	64
5.4	The Use of Historical Information	65
5.5	The Use of Best Practice	65
5.6	Test Structure	66
5.7	Limitations and Error Sources	66
5.8	Future Work	67
5.9	Ethical Considerations	68
6	Conclusions _____	71
	Bibliography _____	73
A	ICD-10 Codes _____	77
B	Test Results _____	79
C	Validation Results _____	81

Preface

This work has been done in collaboration with Lytics Health AB during the spring of 2018. We hope that the project will contribute to the development of a system that can predict if patients with End Stage Renal Disease will be hospitalized within 30 days or not. We also hope that this will lead to better healthcare system since the medical staff hopefully can make better decisions with the support from this system. In this way, the personnel can spend more time with the patients that are in acute need of health, which in the end improves the quality of care. Finally, we hope that the resources at healthcare centers will be used more efficiently and that this system will contribute to a more preventive healthcare system which is more important due to the larger and older population as well as welfare diseases.



Introduction

In countries such as United States, Taiwan, and Japan, end stage renal disease (ESRD) affects more than 1500 people per million in population. End stage renal disease is the medical condition where chronic kidney disease (CKD) has reached stage 5 and the diagnosis is fatal in the absence of kidney transplantation or dialysis [1].

Regular dialysis treatments or kidney transplants are expensive and very exhausting for the patients and their relatives. With a supporting AI system (a system using artificial intelligence) it could be possible to predict which patients who are at high risk of being hospitalized in the near future. The advantage of using such system is to prevent potential hospitalizations or at least be able to prepare or make plans for larger interventions that are needed [14]. The aim is to save money, time and to reduce patient suffering.

Lytics Health AB has developed different decision support systems for patients with ESRD. One of the systems predict hospitalizations among patients with ESRD using patient vital sign time series such as laboratory results, blood values, and habitual data. The predictions of the AI system is relayed to nurses through a dashboard, an email, or a notification. In this way, the nurses can use the predictions together with medical information to support their decisions. This will make the monitoring of patients more efficient, as the nurses easier can identify patients with a high risk of being hospitalized in the near future.

The predictions are given as AI scores which, compared with other patients' AI scores, are an indication of which patients that have the highest risk of being hospitalized within 30 days. If it is possible to find these patients, appropriate interventions could be performed before the hospitalization is needed. This could prevent the development of the disease and is one way to deliver an individualized care. The aim of the AI system is to reduce the number of hospitalizations and otherwise the length of stay in the hospital as it can discover problems earlier and thereby prevent severe illness. It serves as a decision support system for nurses and Lytics Health AB's aim is to constantly explore new data sources and new methods to improve its efficiency.

The main goal of this master's thesis was to examine the importance of the features that are used in a predictive AI system. In order to do this, an in-house version of an existing AI system was reviewed. This version is under development and is not used commercially. The goal of the thesis was further divided into two parts where the first part was to clarify the meaningfulness of the existing features and to find new features that can be used within the system. This was done both by analysing patient data and by applying knowledge from the domain of dialysis

care. This knowledge was obtained by interviewing experts at the dialysis clinics. The information from the healthcare professionals included for example laboratory results, medical measurements, and general routines that are important in order to understand the medical picture of dialysis patients. In the second part of the work, the proposed features were implemented in a test environment, in order to evaluate how the prediction performance was affected by the changes in input data. It was only the performance of the algorithm that was investigated in this project, i.e., the input data to the system and not the classification method itself. Data from Centers for Dialysis Care (CDC) in Ohio, United States, was used during the entire project and the data used in this project consisted of patient information, collected during eight years. The report is further divided into the following parts:

- In Chapter 2 the required background and explanations are given in order to understand the method used in the project.
- In Chapter 3 the reader is introduced to the methods and the important steps that are used to reach the final outcomes of this master's thesis.
- In Chapter 4 the results of this project are presented.
- In Chapter 5 the discussion is interpreting the results as well as future work and limitations of the project.

1.1 Abbreviations

AI	Artificial Intelligence
APD	Automated Peritoneal Dialysis
AUC	Area under the ROC curve
BUN	Blood Urea Nitrogen
CAPD	Continuous Ambulatory Peritoneal Dialysis
CDC	Centers for Dialysis Care
CDDS	Computer-based Decision Support
CKD	Chronic Kidney Disease
CM	Care Manager
CMS	Centers for Medicare and Medicaid services
ED	Emergency Department
EHR	Electronic Health Record
ESCO	ESRD Seamless Care Organization
ESRD	End Stage Renal Disease
FIC	The Family of International Classification
FPR	False Positive Rate
GFR	Glomerular filtration rate
HD	Hemodialysis
HIPAA	Health Insurance Portability and Accountability Act
HP	Hospitalized Patients
IQR	Interquartile Range
NHP	Not Hospitalized Patients
PCP	Primary Care Physician
PD	Peritoneal Dialysis
QOL	Quality of Life
RN	Registered Nurse
ROC	The Receiver operating Characteristic
TPR	True Positive Rate
WHO	World Health Organization

2.1 End Stage Renal Disease

The kidneys' role is to filter the blood and thereby remove waste products from the body system and maintain the fluid balance inside the body. The waste products and affluence fluid are eliminated and preserved in a complex procedure, called glomerular filtration, and are leaving the body through the urine [33]. The kidneys' functions are, except filtering the blood, also to regulate the blood pressure and maintain the balance of the electrolytes and red blood cells within the body. Figure 2.1 shows an illustration of the kidneys together with the ureters and the bladder.

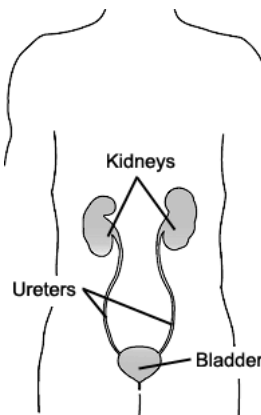


Figure 2.1: Illustration of the kidneys' location inside the body, together with the ureters and the bladder. ¹

Located inside the kidneys is an intertwined group of capillaries, called the glomerulus. The function of the capillaries network is to filter the blood and produce urine consisting of the waste products and redundant fluid. Glomerular filtration is the name of the process whereby the capillaries of the glomerulus filter the blood. In normal cases, the kidneys will receive 20 % of the blood pumped by the heart each minute, and the remaining 80 % of the blood provides the rest of the body with

¹Kidney Illustration [Online image]. (2006). Retrieved May 5, 2018 from http://www.kidneyurology.org/Library/Diabetes_High_Blood_Pressure/Prevent_Diabetes_Problems_Keep_your_kidneys_healthy.php

nutrition and facilitates the gas exchange [26].

A measurement used to check the kidneys capacity, or how much function the kidneys have, is the glomerular filtration rate (GFR) [32]. GFR is an estimation of the passed blood volume through the glomerulus every minute [29]. The medical condition kidney failure can be diagnosed by measuring the level of creatinine in the blood and estimating GFR [47]. For a person with normal kidney function, the creatinine should be approximately 0.5 - 1.2 mg/dL and the GFR should be above 90 mL/min [28, 32].

Abnormalities of the kidneys' function that last for more than three months are explained as chronic kidney disease, CKD. High blood pressure and diabetes are the main causes for CKD and the rest, about one-third of the cases, have other diseases as causes [30]. CKD is classified based on the level of GFR and there are five different stages depending on the severeness of the disease, see Table 2.1 [32].

Table 2.1: Stages of CKD and the respective values of GFR and percent of kidney function for the different stages [32].

Stage	Description	GFR [mL/min]	Kidney function
1	Normal kidney function	>90	~ 100 %
2	Mild reduction of kidney function	60-89	~ 70 %
3	Moderate to mild reduction of kidney function	30-59	~ 50 %
4	Severe reduction of kidney function	15-29	~ 20 %
5	Total kidney failure, ESRD	<15	<15 %

The fifth and last stage is equivalent to ESRD. When a patient has reached ESRD, the kidneys have nearly lost all their ability to do an effective job, which causes severe pain and disorder for the patient. At this stage, the body cannot take care of the waste products and the fluid balance on its own and therefore dialysis or kidney transplantation is essential for survival. It is important to detect and treat CKD at an early stage in order to prevent ESRD.

CKD is a widely spread disease and about 30 million adults in America are diagnosed with this disease [30]. It may occur at any age to anyone, but some are at greater risk to develop CKD and it is proven that patients with diabetes, high blood pressure, a history with CKD in the family and higher age are at greater risk for developing this medical condition. For most patients, it is hard to perceive the symptoms before the CKD has advanced. In the early stages, some of the symptoms can be energy loss, poor appetite, swollen feet and ankles, dry and itchy skin and a greater need to urinate often [30].

2.2 Dialysis

When CKD has progressed to ESRD and the kidney function is about 10 to 15 percent of the normal function the patient must begin with dialysis treatment. Dialysis is a treatment which replaces the function of the kidneys. Two kinds of dialysis are hemodialysis (HD) and peritoneal dialysis (PD) [31]. Approximately two out of three patients with ESRD receive hemodialysis and about one quarter undergoes kidney transplantations and the rest receive peritoneal dialysis [1].

2.2.1 Hemodialysis

A hemodialyzer is an apparatus which works as an artificial kidney outside the body. The patient is connected with the hemodialyzer through an access to the blood vessels. The access can be done in different ways and can be located at different sites on the body. One common access is a so-called fistula which is when an artery and a vein in the underarm is connected together to create a larger entrance to the blood system [31].

The blood is cleaned by a process where the blood is flowing through a semi-permeable filter which is composed of many thin and hollow fibers. Inside the hemodialyzer, waste products and the abundance of salts and fluid are extracted from the blood [35]. The hemodialyzer also keeps the level of different minerals at a certain level inside the body. As the amount of water in the body is adjusted by the hemodialyzer, the blood pressure is controlled during the procedure [31].

Hemodialysis is not a healing treatment for CKD but rather works as an artificial kidney which maintains the balance in the body. The HD treatment is usually done three times a week, but when needed the treatment can be performed more often. Each session takes about four hours but this can also be varied due to how well the kidneys work, the patient body size and how much wastes and extra fluid that the patient must get rid of [31].

Hemodialysis is a complex form of dialysis which demands a vascular access, hence there is a risk of infection. The patients need to visit the clinics several times per week when they use HD, but the advantage with this is that they are under observation by the healthcare clinicians [45].

2.2.2 Peritoneal Dialysis

Peritoneal dialysis is a type of dialysis that cleans the blood inside the body and analogous with hemodialysis replaces the kidneys' functions when the kidneys do not work as they are supposed to do. This is a treatment which usually is performed by the patient themselves in their home. The principle for PD is that the blood will be cleaned inside the body, with help of a dialysate and the peritoneum, which is the membrane that covers the abdominal cavity. The dialysate is brought into the abdominal cavity by a catheter and it is common that the PD patients constant have an amount of dialysate in their abdominal cavity. The peritoneum works as a dialysis membrane and by osmosis, the waste is separated from the blood. The dialysate often contains a substance that contributes and preserves osmosis which maintains the fluid balance in the body [43].

The dialysate has to be lead in and out from the abdominal cavity. This is possible with a permanent catheter of silicone, placed and sewed in the wall of the abdominal. The fluid leaves the body through the catheter after the filtering process

is finished. Continuous Ambulatory Peritoneal Dialysis (CAPD) and Automated Peritoneal Dialysis (APD) are two different types of PD treatments [43].

- CAPD: In this type of PD the patient change the fluid on its own, usually four times each day. The old dialysate with waste products and fluid will be replaced with the new fluid into the abdominal cavity. The time consuming is about 30 minutes each time.
- APD: In APD a machine will change the fluid by night, more often than in the CAPD. This treatment takes about 8-10 hours each night and is often complimented with some CAPD during the day.

In contrast to HD the PD is simpler form of dialysis treatment. Patients with PD have more freedom as they do not need to visit the clinics several times per week as the HD patients do. After a learning period, the patients with PD are therefore able to live a more active and social life [45].

2.3 Comorbidities

A data system with data about patients with renal problematics is available in the United States. The system collects, distribute and analyzes information about CKD and ESRD. From 2005 to 2011, approximately 770 000 adults were initiated with ESRD treatment. About 55 % of them were men and the rest were women [25]. In total, about 10 % of the worldwide population is affected by a disease connected to kidney failure [49].

ESRD is a lifelong, irreversible disease which results in lifelong costs and treatments [19]. In the United States population, about 13 % is affected by CKD. The costs of CKD increase in the same way as the severeness of the disease, which means that stage 5 of CKD is the most expensive stage [18]. In 2009 the total Medicare spending on ESRD in the United States was about 30 billion dollars. This is 23 billion dollars more than it was in 1991, and a lot of improvements can be done to decrease these costs, e.g., identify various aspects of CKD in order to understand the development of the disease [34]. The increased costs can be referred to several comorbidities and the fact that patients become older in today's community [23].

Comorbidities for CKD have for a long time been interesting to study, due to their high costs as one reason [21, 24]. Among the patients with severe CKD, 86 % have one or more comorbidity connected to their original problematic with ESRD. The most frequent reasons for hospitalization among these patients are hemodialysis access complications, sepsis, congestive heart failure, and diabetes. Diabetes is also a risk factor for getting CKD and also increases the likelihood of cardiovascular diseases and death among this group of patients. Other comorbidities that may occur and affect patients with CKD are depression, osteoporosis, and sexual dysfunction as some examples. These comorbidities have been found more common among patients with CKD than for healthy adults [9].

Younger age, female sex, ethnicity, comorbid medical conditions and tobacco use are some factors that are associated with high rates of emergency department visits among patients with ESRD [25]. Most preferably is it to find reasons that can be medicated at an early stage and hopefully by that be able to avoid hospitalization. Infections and cardiovascular abnormalities are examples of comorbidities that can be reduced with right assessments in early stages. Early access to a nephrologist and good placement of fistulas are important factors that are able to improve the

care for patients with CKD and is something that should be in focus. In most cases, improved care also results in decreased costs [37].

Studies have also shown that comorbidities affect the quality of life, QOL, for patients with ESRD. The field of QOL can either be in the physical domain, psychological domain or in the social domain [46]. It has also been proven that the QOL, for patients with CKD, is considerably better if the comorbidities are identified and treatments are initiated according to clinical guidelines. It is therefore of great importance to identify the comorbidities. More research in this area is needed and of high importance to fully understand and confirm the relationships between comorbidities and CKD [9].

2.4 Medical Coding

Medical coding is a standardized method that should describe the entire history of a patient, i.e., which medical events that a certain patient has gone through. It is very important that patients with the same medical conditions can be grouped and compared with patients that suffer from other conditions. The medical coding should allow uniform documentation between different medical facilities. Some of the existing documentation methods are universal and others are country specific. The documentation can be done by using an alphanumeric code system for diagnoses, procedures, medical services and equipment [2]. The medical coding is a type of translation where medical reports from doctors are translated by medical coders according to standardized, predefined methods. Medical coding is not only used for universal understanding for physicians, it is also important for the billing process, where the codes are used to decide the individual fee [2].

There are several medical code systems over the world and in order to create a coding language that can be used all over the world, the World Health Organization (WHO) developed a concept to be able to group different health classifications. This concept is called the Family of International Classification (FIC) and the thought is to improve communication about health. The FIC also made it easier to compare data across different healthcare disciplines.

One of the code sets within FIC is the International classification of Diseases (ICD) and this set consists of different diagnostic codes established by WHO in the 1940s which have been updated several times since then [22]. ICD is used widely across the United States by physicians, nurses, researchers, payers, other healthcare providers and also clearinghouses, which are objective contacts in financial transfers. Health Insurance Portability and Accountability Act, in short HIPAA, is a public law established 1996 in the United States with the intention to improve the effectiveness and efficiency of healthcare systems. HIPAA consists, among others, of standards, unique health identifiers, and code sets [36]. Using the ICD code set is one of the criteria for being in compliance with HIPAA.

The latest revision of the ICD code set is International Classification of Diseases 10th Revision (ICD-10) and this set was introduced in October 2015. Within the ICD-10 code set, the different medical conditions are categorized in order to their class. The length of the ICD-10 codes are from three to seven characters, and could both include a letter, a numeric and any combinations of them. For all codes, the first character always is a letter and the second character is always numeric [17]. The first characters can be seen in Appendix A with belonging description. Collected morbidity and mortality data can systematically be recorded, analyzed and compared

due to the health classification codes. The ICD permits easy storage and retrieval of the data when the diagnoses of diseases and other health problems is translated from words into alphanumeric codes. The codes can then be comprehended by all, talking the same code language [48].

2.5 ESRD Seamless Care Organization

The concept of ESRD Seamless Care Organization (ESCO) was initialized by the Centers for Medicare & Medicaid Services (CMS) in order to coordinate the care around patients with ESRD. An ESCO is characterized by a partnership between a number of healthcare providers in the same geographical area. These healthcare providers are certified by or enrolled within Medicare or just providers of healthcare to patients with ESRD. Their mission is to work together in order to provide a more patient-centered care where the communication between the healthcare providers and the patient is improved [10].

The vision with ESCO is to provide a healthcare system which is more adapted to the patients of today. The healthcare providers, as for example dialysis centers and kidney experts such as nephrologists, are encouraged to come up with new ways of treating and coordinate the provided care. The care is provided to patients with ESRD in order to improve the outcomes for these patients and to lower the costs within the healthcare system. This implies that the ESCOs are accountable for the financial and clinical quality outcomes for the care of the patients [6].

There is a bunch of requirements that need to be fulfilled in order to become an ESCO and one of the requirements is to have at least 500 ESRD patients which are matched to ESCO. To be matched within ESCO, the beneficiary needs to meet a number of requisites. This includes among others receiving dialysis services for more than 120 days, being older than 18 and not having a functioning kidney transplant [7].

2.6 Computer-based Decision Support System

Software that supports decision making is described as Computer-based decision support (CDDS) [44]. They are so-called active knowledge systems, which can be used to perform diagnostics and determine treatment strategies. The system translates information into practice. For example, can large amounts of information within a database can be used by a CDDS system to find connections and draw conclusions, which in turn can be a stable foundation for well-grounded decisions in practice.

Medical information systems were introduced in the 90s with a vision of being able to make decisions that would lead to the best healthcare system. When doctors meet patients they only have minutes to define the problem and to diagnose the patient. They do not have time for advanced research and their statement is therefore based on their current knowledge, and therefore a CDDS system may be of great use [44]. This system can include general information, diagnostic and therapy, communication, reminders etc. The motive for using computer decision system is to give the nursing staff support in their decisions. The thought is not to replace the medical employees, it is rather to improve the effectiveness and knowledge within the healthcare system, and to lower the costs at the same time.

2.6.1 Random Forest

One part of the AI system, which is reviewed in this project, uses the classification method Random forest. This statistical classification method is used to classify data into predefined groups. It is based on the idea that a set of training data is used to find patterns in the data. From the gained observations it can be decided to which group new sets of information belong. Random forest is a classification method which consists of multiple decision trees. The input data to the classifier are features with belonging labels. The features can be of various meaning, medical values as an example. These are described further on in this report. As many other classifiers, decision trees need to be trained or built before it can be used for classification. The training is done using training data which should consist of similar data as the one that later will be classified. The final goal, when the classifier is trained, is to find the right labels, i.e., groups, to the features that are unknown to the classifier.

Decision trees are named after their tree-like structure. When making a decision tree the goal is to find a cut off value for each feature, which as accurate as possible can divide the training set into groups where the classification is correct. In other words, the system tries to create groups where as many subjects as possible have the same label within each group. A decided cut off value creates a branch where the training set is divided into smaller groups. At the next branch, a cut off value is chosen for another feature which separates the training set further. The training set is divided again and again until no further branching is possible. The decision tree is then built and the training session is finished.

A Random forest consists of multiple decision trees which together contribute to the final classification. Before the classifier begins with the training session, some so-called hyperparameters are decided and one of these is the number of trees used within the Random forest. For each tree in the random forest, a subset of input features is randomly selected to split on [4]. Another tree receives another randomly chosen subset with a randomly chosen part of the features. By doing this, trees in different parts of the feature space are created. These trees have different knowledge about the training data and can consequently complement each other which hopefully makes the classification more accurate. When a new set of information will be classified, the information is split by the cut off values in the created trees. Each tree will then do their own evaluation of which label the information belongs to, which results in a vote [4]. The final Random forest classification, which is the collection of votes from all the trees, is therefore based on the results from all the created trees, see Figure 2.2.

Input data, such as medical variables, separately often only contain a small amount of information. The medical information seems therefore as a weak input and it is hard to find a small group of inputs or a single feature that can distinguish between the groups. Single tree classifiers are able to perform a classification with high speed but their accuracy is not that high with this kind of weak input data. A random forest classifier is able to combine the variables through the multiple trees and can, therefore, find the best mixture of how to interpret the variables. In this way, the random forest method keeps the advantages offered by a single decision tree but at the same improves the accuracy [4].

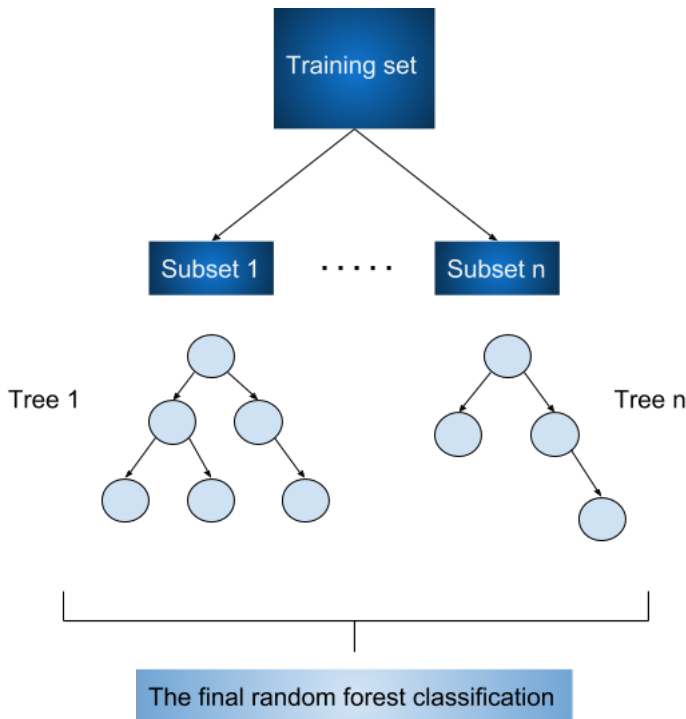


Figure 2.2: Illustration of a random forest classifier. Each light blue node in the trees is a certain feature with a certain cut off value which splits the subset into different parts. The subsets are further split until no more splits can be done. When the bottom of each tree is reached, each tree has made a decision of where to put its vote on. The final random forest classification is a combination of all trees' votes.

2.7 Feature Extraction

Feature extraction consists of two parts: feature construction and feature selection. Data can be represented by a fixed number of features, also called input variables or attributes. A good representation of the data is specific to the current domain and expertise in the field is often needed to build the best representation of the features. The expertise, e.g., a human with great knowledge in the current domain, converts a set of useful features from raw data and this can be done using feature construction methods [13].

2.7.1 Feature Construction

Feature construction is a processing step in order to create a useful set of features. It is a process in which features are built from an original dataset or information within a database. In order to have a high performance of a machine learning method, as for example a random forest, a good set of features is a prerequisite. Feature construction generates a set of powerful features, or a feature space, coming from

transformation of an original given set of input data. The generated feature set can then be used for prediction in machine learning algorithms. Feature construction can be applied when data dimensionality should be reduced or when improvements in the prediction performance are needed. There are many different methods for feature construction but every concept looks quite similar, they consist in general of following steps [41]:

1. Start with a feature space or some kind of initial input information which often is manually collected.
2. Construct a new feature space through a transformation of the original feature space or input information.
3. Select a subset from the newly created feature space.
 - Investigate if the subset of features meet some determined criteria, e.g., if the prediction performance of the machine learning method becomes improved.
 - If the criteria are not met, redo step 3 until the criteria are achieved.
4. When the iteration of step 3 is finished, the new subset of features is seen as the newly constructed feature set.

The initial input information is often consisting of some basic domain knowledge, e.g., medical information from a specific healthcare area, that are manually constructed and determined. As one of the main thoughts with machine learning algorithms is to create connections that usually are difficult to find without the algorithm, it is inefficient to do the rest of the feature construction manually [41]. The way different feature construction methods differ is the methods for doing the transformation, the selection of the subsets and the definition of the predefined criteria in step 3. All the three aspects are important and in different feature construction methods, they can occur in a different order [41].

Features with information about medical history such as clinical variables can sometimes be added together with other information as for examples patient data, weather etc. Adding these features will lead to an increased dimensionality of the patterns and thereby the relevant information may be hidden by possible irrelevant, redundant and noisy features. In order to decide if a feature is relevant or informative feature selection can be used [13].

2.7.2 Feature Selection

The other part of feature extraction is feature selection which is used to select valuable features. Some motivations for feature selection are to increase the algorithm's speed, to limit the storage requirements, to be able to understand the data and to reduce the size of the feature set. Ranking the features according to their individual relevance is one approach to feature selection and there are many methods available for this. There are some limitations with feature ranking such as that less relevant individual features may become relevant together with others. Instead, one can use multivariate methods which also take the feature dependencies into account. A feature will be ranked high if it provides a good separation of the data set [13].

Analogous to feature construction, feature selection also consists of different methods [41]. Which method that is selected is dependent of what is important

within a specific problem. Some methods are fast and independent of the algorithm and some other use the algorithm in order to find the best suitable features. Both statistical methods and cross-validation can be incorporated in to the methods to do the feature selection [20].

2.8 ROC Curve

The receiver operating characteristic (ROC) curve is a graphical plot that can be used to describe a binary classifier's capacity to correctly classify or a diagnostic method's ability to determine whether a patient is sick or healthy [15]. A binary classifier's task is to divide some objects into two distinct groups. The classifier predicts a condition and this condition is then compared with a true known condition. Dependent if the predicted condition is true or false and if the true condition is true or false, different fractions are received. The fractions can be seen in Table 2.2.

Table 2.2: Representation of different fractions from the predicted conditions relatively the true conditions.

	True condition positive	True condition negative
Predicted condition positive	True positive (TP)	False positive (FP)
Predicted condition negative	False negative (FN)	True negative (TN)

From the fractions in Table 2.2, different measurements of the classifier's performance can be calculated. The sensitivity, which is the probability of detection, is described as the true positive rate (TPR) and calculated in Equation 2.1. TPR is, in other words, the amount of correct performed classifications among the number of all the possible classifications that are positive.

$$TPR = \frac{\sum(\text{True positive})}{\sum(\text{True condition positive})} \quad (2.1)$$

The false positive rate (FPR) is the probability of false alarms. FPR is calculated in Equation 2.2 and it describes the number of incorrect classifications made by the classifier among all the negative classifications, i.e., the classifications that should be classified as negative. Specificity is calculated by $(1 - FPR)$ and is the probability that the classifier classifies something as false when the labels say it is false. [27]

$$FPR = \frac{\sum(\text{False positive})}{\sum(\text{True condition negative})} \quad (2.2)$$

In the ROC curve in Figure 2.3, TPR is plotted at the y-axis against FPR at the x-axis. In a binary classification, it is often difficult to separate the objects into two distinct groups or categories[27]. A cut off value between 0 and 1 is determined in order to distinguish between the groups and the ratio between TPR and FPR is dependent on the size of the cut off value. The ROC curve illustrates different ratios at different cut off values.

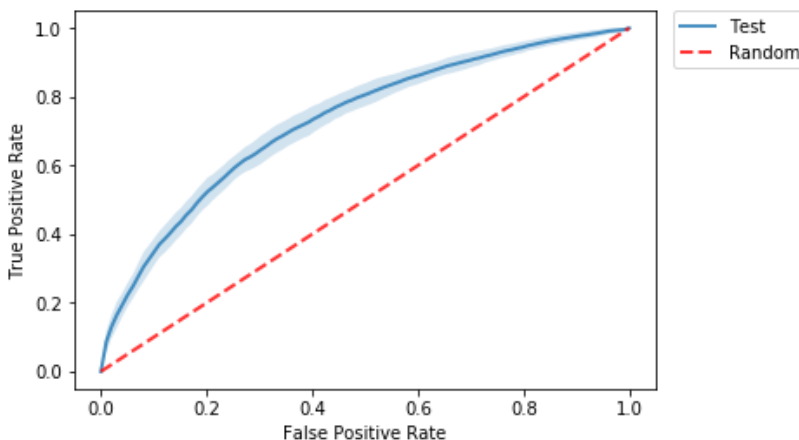


Figure 2.3: The ROC curve from one of the tests performed during the project. The red line marks a test with random classification.

A so-called perfect classification is the one with 100 % sensitivity and 100 % specificity, which is none FN and none FP. This point will end up in the upper left corner of the ROC curve. If the classification is done randomly a line diagonal across the ROC space would be created, from the lower left to the upper right corner marked as the red line in Figure 2.3. Points that are placed above this diagonal are considered to be better than chance while points placed under the diagonal are worse than chance.

ROC can be used to see how well a method, or as in this case a classification, can distinguish right from wrong or in other words separate which patients that will be hospitalized within 30 days from those who will not [16]. One of the most common methods to describe or to interpret the ROC curve is to calculate the area under the ROC curve, AUC in short [16]. In this way, it is possible to receive only one value that can be used to compare different methods, instead of comparing different ROC curves. To summarize, the larger area under the curve, the better is the prediction and the best possible prediction has the AUC value 1.

This master's thesis deals with the analysis of the input data to an in-house version of an AI system provided by Lytics Health AB. In order to be able to do this analysis, the system was first needed to be reviewed in its entirety. In the future, the goals of a system like this are to lower the costs and to use the resources as efficiently as possible. This master's thesis can be seen as a step towards this goal.

The main goal of this project was to investigate the features used in the AI system and to improve the impact of them. In this project, the newly enhanced features were mainly constructed by the use of best practice within the healthcare system. The definition of best practice is well-established procedures and methods accepted by the domain expertise in the current field. These methods are proved to be the most effective and correct. In order to reach the main goal of the thesis in a structured way, the goal was divided into a number of substeps. The substeps, which were drafted in the initiation of the project and investigated through the entire project, can be seen below:

1. **Identification:** Identification of the existing system and the test environment; how it works, the input data and the output data.
2. **Best practice:** Collect information about dialysis work and healthcare system in general, by using domain expertise.
3. **The features:** Identification of the existing features, preprocessing and engineering of new features.
4. **Test engineering:** Generate tests with different sets or combinations of features.
5. **Validation:** Validate the results of the tests.

As the last part of the project, the potential of the results was analyzed whether they could improve the prediction performance of the AI system. The following part of this chapter describes the work that has been done, including assumptions that have been made and the steps that took the project forward.

3.1 Identification

Identification of the AI system was an important part of this project, as it provided the necessary background to understand the system. This was made during the first part of the project and gave a solid base to build the rest of the project on.

Lytics Health AB is a business associate to Centers for Dialysis Care (CDC) Incorporated and all data that was used in this project was provided by CDC. Centers for Dialysis Care is a non-profit medical provider of dialysis treatment and medicate patients with kidney failures. The center is located in the north of Ohio and except from taking care of patients with renal failures they put a large part of their work to educate and improve the care for patients with ESRD [5].

The available database consists of information about over 7000 unique patients collected during eight years. Since the patients are visiting the clinics several times a week due to their dialysis treatments there were lots of information to base this master's thesis on. The information that was available for the AI system originates from primarily four different sources, see Figure 3.1 below. At CDC they use an electronic health record (EHR) to save and store patient information. In addition to this medical information, information is also collected from Care Managers (CMs) and dialysis machines. When a patient receives a dialysis treatment, the dialysis machine records different values such as dialysate temperature, liters processed or fluid removed. The source Statewide Health Information Exchange is representing the information from other hospitals, clinics or centers where the patients receive care. The information from the explained sources is processed and transferred into the AI system. The existing AI system uses this information and the results are presented on a computer dashboard which is the graphical user interface, GUI.

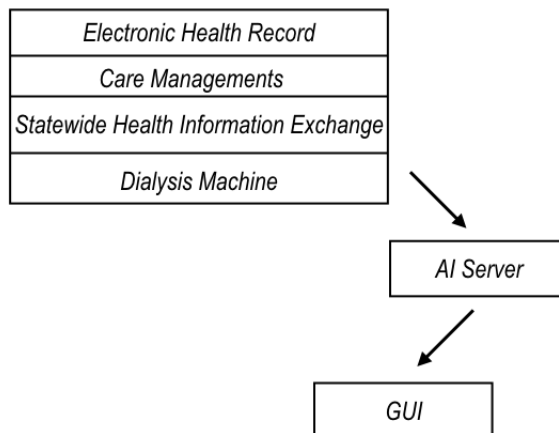


Figure 3.1: There are four different sources from where the information used within the reviewed AI system originates. The information is saved on a server before it is processed. The results from the system are then presented to the clinicians in the GUI.

In addition to the planned dialysis treatments, there is four kinds of visit types that the patients can do when they are in need of acute care. These four different ways of seeking care contribute with information stored in the database. The visits are named and defined below.

- **Emergency Department Visit**

These are the patients that go to the emergency department (ED) and are discharged from the ED when they have received the needed care. The length of stay is not longer than 48 hours in these events.

- **Emergency Department Visit Resulted in Admission**

If the patients become admitted after the visit to the ED, the EMR should automatically update the event from Emergency Department Visit to Emergency Department Visit Resulted in Admission.

- **Hospitalization**

This is an event where the patients are directly admitted to the hospital without a visit to the ED.

- **Observation Admission**

This is a special form of hospitalization that not all hospitals provide. Observation admissions are often located close to the ED and can be a 23, 48 or 72 hour event.

All the patients in the database have ESRD and these are the patients that are included in the training data for the AI system. It is only the patients within ESCO that get hospitalization predictions from the AI system. The database consists of several tables with different information about the patients. The information is overlapping in some cases and data is missing in others.

To identify the patient group that will be used in the project, general information about them were collected. The information of interest included both common diseases for the patients, distribution of the different healthcare visits and also information about these patients such as age, gender, months in dialysis etc.

3.1.1 The AI System

The core of the reviewed AI system is a random forest classifier which uses a number of selected features to predict whether a dialysis patient will be hospitalized within 30 days or not. The features in the algorithm are medical data from the patient's previous visits at the clinics.

The intended users of the system are professionals working at dialysis clinics. These are, primarily, CMs which are responsible for ESCO patients. The purpose of the final prediction algorithm is to identify the patients having the most need of care. In this way, the nurses get an early indication of which patient that may be hospitalized in the near future. Due to this they can initiate actions and provide the correct treatments before the patient reaches a severe health state. The main thought with the AI system is to aim for a preventive care rather than a treatment care.

3.1.1.1 Input Data

The reviewed AI system gets information, also called input data, which is processed from the sources in Figure 3.1. The processed data results in several features containing information about the patients, collected from the database. The information in the database is structured in different tables which describe different aspects of the patients' medical histories.

As the primary sources for the input data, four tables from the database are used and they consist of the following information:

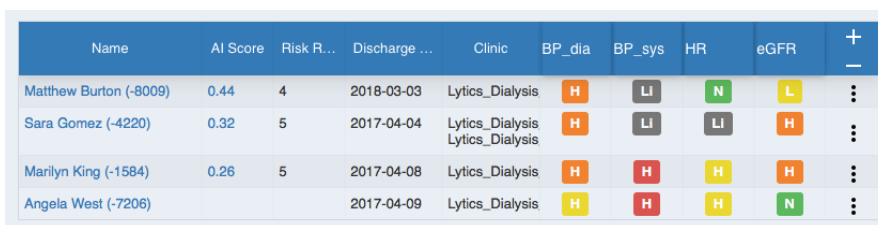
- Administered medication information
- Information about dialysis treatments
- Patient measurements and laboratory results
- Information about the patients' diagnoses and completed forms

Information from these tables are processed into features used as input for the algorithm. The features vary every time the system runs and this depends on the available data that the system is training on and the used feature selection method.

The feature values are summarized weekly, due to the structure of the system. The input data consists of several rows where the patients have personal rows with information from each week. The training data consists of all data rows that have been collected up to 30 days before the day of prediction. A larger amount of training data hopefully gives the algorithm a broader knowledge about the patient group and their complaints.

3.1.1.2 Output Data

The output data from the system are predictive AI scores. The AI score is a number between 0 and 1. The AI scores are the final results from the classifier and cannot be used directly by the CMs. AI scores from different patients are compared in order of magnitudes. The patients at each day with the highest AI scores are considered to the highest risk of being hospitalized within 30 days. Figure 3.2 shows an example of the output data that the intended users can see on their screens.



Name	AI Score	Risk R...	Discharge ...	Clinic	BP_dia	BP_sys	HR	eGFR	+
Matthew Burton (-8009)	0.44	4	2018-03-03	Lytics_Dialysis	H	LI	N	L	⋮
Sara Gomez (-4220)	0.32	5	2017-04-04	Lytics_Dialysis Lytics_Dialysis	H	LI	LI	H	⋮
Marilyn King (-1584)	0.26	5	2017-04-08	Lytics_Dialysis	H	H	H	H	⋮
Angela West (-7206)			2017-04-09	Lytics_Dialysis	H	H	H	N	⋮

Figure 3.2: The interface of the reviewed AI system, with imaginative patients.

3.1.1.3 Missing Data

Missing data is not unusual when working with information in databases and this database was not an exception [11]. There were both cases when data was completely missing and cases when the information within the tables was contradictory which made the data unreliable. The data was mostly considered to be usable, but if all rows with missing or unreliable data were removed the size of the data would decrease a lot.

There are many ways to handle missing data and in the AI system, this is handled in various ways. One can calculate a mean value for all patients' values according to a specific feature. This mean value can be used when a patient is missing this particular piece of information within the database. Calculate the mean or median value for the patient's previous values is another way of estimating missing information. Forward filling, when the last existing value is reused when data is missing, can also be used. A backward filling is the opposite of a forward filling and uses the next occurring value to fill in information backward in time. All of these methods were in this projects used when input data were preprocessed as well as just adding zeros when a value was missing.

It was quite common that a patient could be listed as hospitalized at the same time as she or he was listed as discharged. This made the data contradictory and uncertainties arose whether a patient was visiting the emergency department, was hospitalized or discharged. A lot of time and efforts has been spend on creating hospitalization data with no overlaps in information and time. To be able to do this, assumptions were needed to be done before the classifier could train and then make the predictions.

3.1.2 Test Environment

A version of an in-house AI system has been used during this project. This version is a working prototype and all the tests within this project have been done in a test environment separated from this reviewed system. The test environment was originally created within another master's thesis, but since no major changes were needed in order to add new tests to the environment, it was decided to be used in this project.

The test environment was created as an extraction from the AI system. The system is built in a level like structure with multiple methods with different dependencies. In other words, the methods are depending on each other and some of the methods cannot be executed without the results from other methods. To create a test environment as similar as the reviewed AI system as possible, a delimited part of the system was identified and separated from the system.

As the test system is built, it uses a file with 306 different features in total. These features will represent the feature matrix called Baseline and were during this project grouped into four different groups based on what they describe. The groups can be seen in Table 3.1 and their content is described further on in this report.

Table 3.1: Result after grouping the features into four groups. The proportion indicates the size of the groups, relative each other.

Group name	Number of features	Proportion of total group size
Medication	55	18%
Dialysis	24	7.8 %
Times since healthcare event	100	32.7 %
Laboratory results and diagnoses	127	41.5%

To perform a test in this project, the input to the test environment was changed and then the system did the classification in the same way as the reviewed AI system. The change of the input can be done in multiple ways, for example, can new features be added to the Baseline, the Baseline can be modified or existing features within the Baseline can be removed.

When the features used as input data were decided, the prediction was executed at 48 predefined days between April 2013 and November 2016. At each prediction day, the training of the classifier and the predictions were performed in the same way as in the existing AI system. This was done by doing the training session at all the information about the patients collected up to 30 days before the day where the prediction was done. The predictions were performed at the patients that are qualified for the reviewed AI system. These patients fulfill a number of criteria, for example, that they are active ESCO patients and are not hospitalized for the day of prediction.

3.2 Best Practice

One common way to create relevant and useful features is to use domain expertise. This was something that not has been widely adopted in the system before. When this project was in its initiation phase it was decided that the new features should be generated with best practice within the healthcare system in mind. The definition of best practice was in this project the use of healthcare expertise. This meant that the features were created after how the healthcare system works today with its policies and guidelines due to dialysis care. A lot of information was needed to be collected in order to incorporate domain expertise in the feature processing steps. Except reading relevant literature, information was gathered by discussions and interviews with experienced persons.

During the entire project, many discussions and dialogs with an Application Specialist at Lytics Health AB have been held. This person has long experience as a nurse and has contributed with knowledge about, among others, the structure of the healthcare system and different aspects when nursing patients with ESRD and chronic diseases in general. In addition to this information source, questions were sent to CDC as a way to understand the daily work at a dialysis center and how patients are treated.

In this way, the variables and overview assessments used in best practice were identified and this was something that was of high importance when the new features were processed. Twelve questions were asked by email to CDC and they were structured into three groups:

1. **Dialysis patients and the daily practice**
2. **Comorbidities, tests, and routines**
3. **Evaluation of patients and the future**

The different groups of questions were chosen in order to understand how the dialysis care works today. Therefore they are quite general and the thought was to give the interviewees freedom to describe the nursing of dialysis patients without too much guidance from the questions. The questions according to the three groups are shown in Table 3.2, 3.3 and 3.4.

Table 3.2: Dialysis patients and the daily practice.

Question
1. Can you tell us about your daily work? For example, how many are you in the staff? How many patients do you have? How many doctors? How often do the patients receive care? How many patients per nurse?
2. As a patient at your clinic, do one have a special contact to a nurse or some other clinicians? Someone that one as a patient can turn to if something happens or if one needs to talk about something? Or a person that follow up between the dialysis treatments?

Table 3.3: Comorbidities, tests and routines.

Question
1. What is the first thing you look at when a patient visits you? Overall impression? Tests? Laboratory results?
2. Common diseases due to the kidney failure? How do you notice them? What do you base your evaluations on?
3. Do the hospitalized patients have different phenotypes? Does it exist one or more “typical” patients that get more sick than other?
4. Are there some common tests that the patients go through? E.g blood pressure, blood values etc? What are you looking at/interpreting from the tests? Do you have some standardized tests that are performed at each dialysis treatment and/or regularly?
5. In which order is everything done? What is the first common indication that something deviates from the normal? Different blood analysis? Medications? Dialysis?
6. How does the routines look like when a patient is close to its first dialysis treatment? Formulary? Tests? Does patients redo this after a while?
7. Are there other variables that you have seen affects the patient’s general conditions? If so, how do you notice/measure this?

Table 3.4: Evaluation of patients and the future.

Question
1. Approximately how long time do you have to evaluate a patients’ condition? I.e how long time do you have before you have to draw conclusions about the patient’s health?
2. Can you see differences in the patients’ outcomes, depending on who is evaluating the patients’ conditions, with regard to experienced nurses or new more green ones? If so, how do you handle this?
3. How do you think the future’s dialysis care will look like? What would be needed to make it even better? For the patients? For the clinicians? For the relatives?

The results from the different questions are presented in Section 4.2. The meaning of the questions was to understand the working processes that are used in the healthcare system, both regarding the treatments and diagnostics. With knowledge about this, the feature engineering could hopefully be done more efficient, realistic and with more feeling for the patients and the healthcare system. Another interesting aspect of this was to see if the use of domain expertise in feature construction will improve the existing AI system.

3.3 The Features

Feature selection for an AI algorithm plays a major role in the prediction performance [42]. In this project, feature selection was done by hand with the intention to create the most advantageous set of features. The selection was done with best practice in mind. The next step was to construct the features, both with respect to their content and shape. The content of each feature set deviates from each other, but their shape must be the same in order to combine the different sets. The shape was particularly important when the created feature matrix was utilized in the test environment. All of the feature sets are described by a feature matrix with multi-indexes, made of patient identification numbers and dates. Each column in the matrix are named as the features and contains their values, see Table 3.5.

Table 3.5: Illustration of the shape of the feature matrix, where the imaginative Time and Patient ID together create the multi-indexing.

Time	Patient ID	Feature 1	...	Feature n
2013-04-07	8009	0		1
	4220	2.34		0
	1582	3		0.52
⋮	⋮	⋮		⋮
2016-11-13	8009	2.15		1.73
	1582	1		0.89

When new features were created in this project a standardized working process was determined and used. The method that has been used can be explained by the different steps in Figure 3.3. First of all feature selection was done and the inspiration and ideas to the selection came both from healthcare expertise and from analyzing the already existing features. New features were both made from brand new information from the tables within the database and from old features that became modified in another way.

The information that was needed in order to build the new features, was collected from the database. The database consists of several tables with information about patient data, medical data, and hospitalization data. The content within the database limited the possibilities of creating all kinds of new features. The first step was to identify tables with right and useful information. Thereafter the information needed to be preprocessed in order to create the desired features. The preprocessing step varied in complexity whether the content of the new features differed much from the original raw information in the database or not. Some common preprocessing steps was to handle missing data, calculating mean values or differences.

The last step in the feature construction part was the feature engineering, which was about to get the right shape of the feature matrices. The shape of the matrices was the same as for Baseline and this was important in order to merge the different set of features together. The feature engineering step was the same for all new sets of features. The most important part of this step was the work when the daily information was translated into intervals of weeks. The new weekly dates were in the number of 48 and used as multi-indexes together with patient identification numbers.

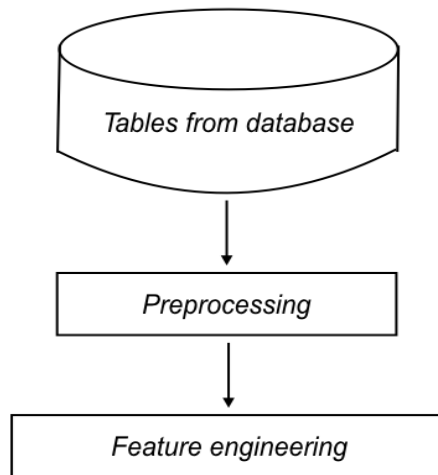


Figure 3.3: Different steps in the method for feature construction.

Feature construction was in this project done within all the four feature groups; Medication, Dialysis, Time since healthcare event and Laboratory results and diagnoses. The steps for the feature construction varied for the different features, especially in the preprocessing part. The production of each feature group is therefore described below according to the feature construction steps and resulted in one or more sets of new features. The definition of new features is in this project the features that have been constructed during this project if nothing else is specified. The newly constructed features consist of both completely new information for the system as well as old information stored or processed in a new way.

3.3.1 Feature Importance Method

In order to compare the different sets of features, important features used by the system were identified by the built-in method *feature_importances_* [40]. As said in Section 2.7.2, there are several ways of selecting which features that are useful for the system in order to do correct predictions. In this project, the built-in method was chosen to identify the most important features as this was available in the test environment.

The feature importance is in this method defined as the Gini importance. When the training data in each tree is divided by the nodes into different subsets, see Figure 2.2, the aim is to create two homogeneous groups where almost all patients are hospitalized in one group and not hospitalized in the other. The Gini impurity is a measurement of how well this separation is performed. For each node, the Gini impurity is decreased [38]. The Gini importance is then calculated by the total decrease in impurity divided by the number of trees. In this way, all the trees do

their own estimation of the feature's importance, and the final result is a summation over how well each feature can separate the training data into the two homogeneous groups[12].

The feature importance method was used after each prediction session. The features were given an importance score after how the reviewed AI system interpreted the influences of all specific features. The features were then ranked regarded to their feature importance score and the higher up a feature was placed the better this feature was. By looking at the ranking, conclusions could be drawn whether the features were considered by the system to be helpful in the prediction or not. This measurement gave a guidance of which features that were important for the system in order to perform in its best way, i.e., to divide the patient into the two groups after the two labels, hospitalized and not hospitalized.

3.3.2 Medication

The existing medication features from the medication group describes which medications that are given to individual patients. If a medicine is given more than once a week, multiple features are created for this specific medicine. As a result of this, the total number of medication features is about twice as large as the actual number of medicines which contribute with medical information used within the system today.

In the database, there are almost 20 000 medicines and each of them belongs to one or more medication groups. The number of medication groups are about 40 times smaller than the number of medicines. When the groups were studied closer, it could be seen that the groups in many cases had the same meaning or even in some cases the same names. Therefore a large portion of effort was laid to regroup the medicines. The regrouping resulted in 16 new and unique medication groups, which was almost 98 % fewer groups than it was before. In the existing system only a few of the different medicines were used, and from these, the medication features were created. The value of the old medication features was the amount of taken medicine at a special occasion.

Two tables consisting medication information could be found in the database, one with information about prescribed medicines and one about the administered medicines. The difference between these is that the administered table only gives information about the medicines that are given at clinics by nurses. The table with the prescribed medicines includes information about all medicines that the patients have a need for, except for the acutely given medicines which are in the administered table.

Today the reviewed AI system only use information about the administered medicines at the clinics. The system does not consider the prescribed medicines, which is a big part because the patients are lifelong medicated since they got ESRD, and therefore take a lot of medicines at home. The mainly administered medicines at clinics are intravenous doses. There is a lot of unknown information about the doses the patients take on their own. There are also a lot of exceptional cases about medications, different administration methods, dose volumes, intervals etc. Therefore it was decided during this project to identify a reasonable number of medicine groups where the medicines could be divided into. By doing this the thought was to incorporate all kinds of medicines into the system.

3.3.2.1 Administered Medicines

The preprocessing steps in the feature extracting from the table with the administered medication were roughly done in three steps. The first step was to identify the current medication groups and their attached medications. The next step was to reduce the number of groups and merging groups with the same content together. In this step, the medications were categorized into right medication group. The unique medication groups were found through discussion with an expert in the domain of healthcare system. As the last and most demanding step all information from the given administered medication data were used together as they were divided into the new medication groups.

Since the existing test system runs in weekly intervals the new features were needed to be constructed in the same way. To handle the processing with the different dates the information was collected and summerized during predefined intervals. This was done by converting all dates to the end date of the closest interval. The meaning of this feature engineering step was to get a table consisting all information about the administered medication given within the intervals for all the patients. This newly created table consists of information about how many medicines from each medication group the patients have received during a defined interval.

The constructed feature matrix consists of 16 medication groups. Every column of the medication feature matrix tells if the patients have had an administered medication in the specific group or not. The value is the number of medicines from a group during a specific week interval, and could, therefore, be zero or higher. This set with features is in this report called MedAdm.

From the database

- Information about administered medicines
- Information about medication used at the dialysis center

Preprocessing of the administered medicine features

- Identification of medication groups
- Reduce the number of medication groups
- Divide all administered medicines within the groups

The final feature set from the feature engineering

- MedAdm (16 features in total)

3.3.2.2 Prescribed Medicines

In the same way as above, the prescribed medications were categorized into medication groups. Every prescribed medication had a start date and, in most cases, an end date. When end date was missing, it was set to the current date the code was executed. The preprocessing steps for these features were quite similar to the steps for the administered medicines, but in this case, the medicines were given in intervals and not on specific dates. This was something that differed from the administered medicines and resulted in one more step in the preprocessing part of the feature extraction. This was done by creating intervals of days between the start day and the end day. The decided way to handle the different dose frequencies during an interval was to just analyze whether a medication would be taken or not during a specific week.

In this case the constructed feature matrix also consists of 16 medication groups, as the shape of the administered feature matrix. Every column of medication groups tells if the patients have a prescribed medication in the specific medication group or not. The value is depending on the number of medicines from a group during a specific week and could be zero or higher. Henceforth in this report, this set with features is called MedPre.

From the database

- Information about prescribed medicines
- Information about medication used at the dialysis center

Preprocessing of the prescribed medicine features

- Identification of medication groups
- Reduce the number of medication groups
- Divide all prescribed medicines into the groups

The final feature set from the feature engineering

- MedPre (16 features in total)

3.3.2.3 Merged Medication Features

The two new sets of medication features were also modified in other ways. They were merged into one set with medication features by simply being added together. As a way to introduce historical changes in the patients' medications, the method rolling mean was applied on the medication features. This was done by calculating the rolling mean values of medicines during four-week intervals. From this feature engineering four new sets with features were formed; MedMerged, MedAdmRolling, MedPreRolling and MedMergedRolling.

From the database

- Information about administered medicines preprocessed to MedAdm
- Information about prescribed medicines preprocessed to MedPre

Preprocessing of the new medicine features

- Merge the two medication feature sets into one new
- Calculate the rolling mean of the administered medications
- Calculate the rolling mean of the prescribed medications
- Calculate the rolling mean of the merged medications

The final feature sets from the feature engineering

- MedMerged (16 features in total)
- MedAdmRolling (16 features in total)
- MedPreRolling (16 features in total)
- MedMergedRolling (16 features in total)

3.3.3 Dialysis

Dialysis is the feature group where all the measured values connected to the dialysis treatments are collected. Information in this group can, for example, be how many liters blood that was processed in the dialysis session or the amount of fluid that was removed from the patient during the session. Features for the pulse and blood pressure tests, describing the patient's pulse and blood pressure in different positions, are also included in the existing dialysis feature group. In this part of the project, the focus was on creating features according to the blood pressure, the blood flow rate and the weight of the patient after a dialysis treatment.

3.3.3.1 Blood Pressure

Today the input data within the dialysis group consists of many different variables describing the blood pressure. Blood pressures are measured in sitting, standing and lying position, both before and after the treatments. It can also be measured during the treatment. As the reviewed system works today, features are created from all of these positions except the lying position. Often a patient only got values on one of these features connected to a specific treatment, depending on the position when measuring. Because, if a measurement is taken in the sitting position it is usually not taken in the standing position also, and vice versa. This results in a lot of missing data and is today handled by filling in the empty rows with a predefined blood pressure value for all patients. The features used today consist of both start and end values for the different position measurements as well as a systolic and a diastolic part.

The sitting position was the most common approach for blood pressure measuring,

according to the database, and the second position was the standing position and these two make the majority. Therefore the new features were made from these positions. The first approach to handle the missing data in different positions was to establish a criterium with help of best practice. The criterium that was investigated is:

“First we look at the sitting position value, and if it does not exist we look at the standing position value. If that value also is empty we use value X”.

This was something that could reduce the number of blood pressure features. By doing this it would be a declination when comparing the values, due to the different positions. Therefore a new rule for handling the missing blood pressure values was needed to be created. This new rule became to keep the values for sitting and standing position as two separated feature sets. In cases where no values for measurements existed, another personal blood pressure value for the actual position was used. As the blood pressure can differ between individuals, it was decided to only use the measurement from the same patient when filling missing data.

The filling of the empty positions was done by replacing them with the last occurred value, measured at the same position as the missed value. If no last value was found the next value closest to the specific date was used instead. In this way, the information about the blood pressure only consisted of personal values. If the patient has a start value at some position, but no end value, the end value was set to the start value. The same happened if the start value was missing, it was replaced with the end value. In cases where the patients lack all values, everything was set to 10 000. This because no useful information was available and this was indicated with 10 000, since this is out of range for possible blood pressure values.

Both the start value and end value were divided into two parts, one part for the systolic pressure and one for the diastolic pressure. To be able to see if there was a difference between the start and end value, and how big it was, the differences of the start and end values for the systolic pressure and for the diastolic pressure were calculated. This was calculated as a percent difference, explained in Equation 3.1.

$$\frac{StartValue - EndValue}{StartValue} * 100 \quad (3.1)$$

The method rolling mean was used as a try to include medical history about the blood pressures. These rolling mean features were constructed in a three months interval, which means that values three months earlier were summarized and divided by the number of weeks during that time.

The new blood pressure features gave both new information to the system and stored old information in new ways. The blood pressure feature sets are in this project called BPsitting, BPstanding, and BPRolling. The first two consists both of the blood pressure values and the difference values. The last one consists of only the rolling mean values, both for sitting position and standing position, and their start, end, systolic and diastolic values.

From the database

- Information about the sitting blood pressure
- Information about the standing blood pressure

Preprocessing of the blood pressure features

- Empty positions are replaced with last occurred values
- If no earlier value exists, the next value closest to the current day is used
- End value is set to start value if no end value exists, and the opposite when no start value exists
- Value is set to 10 000 if neither start or end value is available
- Calculate the difference between start and end value
- Calculate rolling mean in three month intervals

The final feature sets from the feature engineering

- BPsitting (6 features in total)
- BPstanding (6 features in total)
- BPRolling (12 features in total)

3.3.3.2 Blood Flow Rate

The blood flow rate is an important parameter for dialysis patients, and determine how fast the blood flows through the dialysis access. In the reviewed AI system a feature connected to the average blood flow rate was used, but this value cannot by itself describe whether it is normal or not. This because the blood flow rate is personal for each patient. In a database table, information about the prescribed blood flow rate existed. This information was used in the new features to see if the blood flow rate followed the recommendations or not.

The first preprocessing step was to fill in missing data. For each patient, there was a prescribed value for the blood flow rate valid for a specific time interval. Each prescription consisted of a start value and an end value for the blood flow rate. The prescriptions were regularly updated or changed. When no end date was determined, it was set to the specific day the classifier was training on.

The next important step was to connect the prescribed values with the measured values from the treatments, which were the average blood flow rates. This was done according to the patient identification number and the current date. To do this, the session dates were connected to the prescribed value intervals, set by the clinicians. The difference was then calculated by taking the average blood flow rate subtracted with the prescribed blood flow rate, see Equation 3.2. Since the dates must be connected to weeks, and several events occurred during a week, the mean value of all differences during a week was calculated.

$$BFDiff = BloodFlow_{\text{Prescribed}} - BloodFlow_{\text{Measured}} \quad (3.2)$$

To include history, the rolling mean method was applied on the blood flow rate differences. The rolling mean was calculated in three-month intervals. The created feature sets with the blood flow rate features are named BFDiff and BFRolling in this project.

From the database

- Information about the average blood flow rate at dialysis treatments
- Information about the prescribed blood flow rate

Preprocessing of the blood flow rate features

- Insert start date and end date and fix intervals
- Calculate the difference between the average blood flow rate and the prescribed blood flow rate
- Connect dates to weeks by summarize events and take the mean of their values
- Calculate rolling mean values in three month intervals

The final feature sets results from the feature engineering

- BFDiff (1 feature in total)
- BFRolling (1 feature in total)

3.3.3.3 Post Weight Difference

Dialysis patients measure their weight before and after a treatment. The weight after a dialysis session is compared with a dry weight which is the calculated goal weight for each patient. This comparison is a good indication if enough fluid was removed from the body during a dialysis treatment. For a dialysis patient, a lot of fluid is needed to be removed as they collect body fluid due to kidney failure.

Information about the post weight and the prescribed dry weight were preprocessed in similar ways as the blood flow rate features. There were prescribed values within different time intervals. The prescriptions were regularly updated or changed. When no end date was determined, it was set to the specific day the classifier was training on. The treatment dates were connected with the right intervals in order to compare the prescribed values with the measured weights after the dialysis treatments. The difference between the post weight and the prescribed post weight was determined by taken the measured value subtracted from the prescribed value, see Equation 3.3.

$$WDiff = Weight_{\text{Prescribed}} - Weight_{\text{Measured}} \quad (3.3)$$

The rolling mean value for three months was calculated to include historical data. Since the dates must be connected to weeks, and several events occurred during each week, the mean value of all differences during a week was calculated. The feature sets constructed from post weight information are named WDiff and WRolling.

From the database

- Information about the weight after a dialysis treatment
- Information about the prescribed dry weight

Preprocessing of the post weight features

- Insert start date and end date and fix intervals
- Calculate the difference between the measured post weight and the prescribed dry weight
- Connect dates to weeks by summarize events and take the mean of their values
- Calculate the rolling mean values in three month intervals

The final feature sets from the feature engineering

- WDiff (1 feature in total)
- WRolling (1 feature in total)

3.3.4 Time Since Healthcare Event

This group of features describes when the patients last visited the hospital or the emergency department respectively. In the reviewed AI system, each patient has 50 features where it is possible to store how many days it has been since the last 50 hospitalizations events. There are also 50 features dedicated to when the 50 last emergency visits were done by the same patient. Even though there are many features that describe the time since a healthcare event occurred, there is no patient that have as many healthcare event to be able to use all 100 features.

Today these features engage approximately 30 % of all the features that the reviewed AI system builds its decision trees on. It felt redundant to have a lot of features that were not used by the majority of the patients. Therefore the work with these features was primarily focusing on replacing these features with fewer features that contained the same or preferably more information. As a first step towards an improvement of these features, three new features were created. How these new features are developed is described below and together they constitute the new feature set named TimeSince.

3.3.4.1 Number of Admissions

The two first features created as an attempt to replace the old time since features gave information about the number of admissions during a predefined time period for each patient. It was decided to divide this new information into two categories. The first category was to count all hospitalization admissions during a time period. Within this category, the events called Hospitalization and Emergency Department Visit Resulted in Admission were summarized. The other category contained only the event Emergency Department Visit. The time interval was set to four months as this was considered to be a reasonable time range according to domain expertise. The creation of these features was quite simple, once the dividing into the two categories was done and the time interval was defined. This feature construction was done by calculating the total number of admissions during the last four months in each of the two groups, at each day of interest. The calculated numbers were then used as features in the TimeSince set, together with the other new time since healthcare event feature described below.

3.3.4.2 Hospitalizations per Week

As a complement to the number of admissions during the previous four months, the ratio between the total number of admissions and weeks in dialysis was calculated and used as a feature. In order to do this, the number of weeks in dialysis and the total number of admissions was connected to each patient. Since these values were changing over time, the feature values were recalculated during each prediction. The feature engineering for this feature was the same as for the other features in the TimeSince set.

3.3.4.3 Top Three Time Since Healthcare Event

Besides the three new features which together were the feature set called TimeSince, an array was done where only some of the old time since healthcare event features were kept. By studying the feature importance the conclusion was drawn that the system favors the time since healthcare event features over many of the other features.

As an attempt to reduce the number of time since healthcare event features, it was decided to keep only six features of these. The first three features describe the number of days since the last hospitalization, the second latest hospitalization, and the third latest hospitalization. The last three features were defined as the number of days since the last emergency department visit, the second latest emergency department visit, and the third latest emergency department visit. With best practice in mind, the choice of the three latest events from each category was considered to be reasonable and of interest for the intended user of the AI system. This set of features is called TimeSinceTop3 in this project.

From the database

- Information about the patients
- Information about hospitalization events
- Information about emergency department events

Preprocessing of the new time since healthcare event features

- Calculate the number of admissions, divided into two groups
- Calculate the ratio between hospitalizations per weeks
- Save the top three existing time since healthcare event features

The final feature sets from the feature engineering

- TimeSince (3 features in total)
- TimeSinceTop3 (6 features in total)

3.3.5 Laboratory Results and Diagnoses

The largest group of the four determined feature groups is the group which mostly describes laboratory results and diagnoses. Half of this group, almost 30 features describing the diagnoses, i.e., different ICD-10 codes. Laboratory results consist of different tests performed by the laboratory connected to the clinics and most of these are blood tests. The blood tests vary a lot, some of them are common analysis such as mineral content within the blood and some others are more specific for dialysis patients as for example blood urea nitrogen (BUN). Features describing patient data, such as the height and weight of the patients, are also included in this group of features.

3.3.5.1 Diagnoses

Features connected to diagnoses were made from the table in the database that contained hospitalization information. The first character in the ICD-10 code represents a specific group of diseases, see Appendix A. The way these features were preprocessed, was that every diagnose was sorted into diagnoses groups together, according to their first character.

The date when the diagnosis was established was determined to be the patient's admitted date. These dates were identified and then placed into correct week interval. Diagnoses with same characters and admission weeks were grouped together. The value of the feature is the number of diagnoses during a week, and if there are no diagnoses the value was set to zero. In this way the features will tell when the diagnosis was established, to which diagnosis group the diagnosis belong to and how many diagnoses the patient has had during the specific week.

In order to investigate the importance of the new diagnosis features, it was decided to create three sets with features within this category. The first category contained all new diagnosis features and this set is called ICDAll. As an attempt to study different aspects of the diagnosis features, the ICDAll was divided into two parts. The first set was called ICDHosp and contained the diagnoses received by the patients' when they became hospitalized. The second part carried information about the diagnoses received by the patients' at the emergency department and is named ICDER. In addition to these three feature sets, the rolling mean method was applied on the ICDAll set. The resulted in a new feature set, called ICDAllRolling, which describes the diagnoses a patient has had during the last three months.

From the database

- Information about hospitalization events

Preprocessing of the diagnosis features

- Identification of all ICD-10 codes
- Identification of admission dates
- Count diagnosis in each group

The final feature sets from the feature engineering

- ICDAll (23 features in total)
- ICDHosp (23 features in total)
- ICDER (23 features in total)
- ICDAllRolling (46 features in total)

3.3.5.2 Age

As an attempt of improving the predictions, the patients' ages were added as a set of features. Since the Baseline's features have multi-index based on dates and patient identification numbers, the age was required to be added to each row. To do that the age was determined for all patients at the predefined classification dates. The feature set connected to the age is in this report be called Age.

From the database

- Information about the patients

Preprocessing of the age features

- Connect patient identification number to age

The final feature set from the feature engineering

- Age (1 feature in total)

3.3.6 Removing Features

When analyzing the content of Baseline the feature values were of great interest to study. Trying to understand the features and their values was a large part, according to understand the input of the reviewed AI system. Some of the features in Baseline were difficult to define and understand, due to strange names and values. Other features consisted of one single value, for every patient and date. This could not contribute with new information to the system as no values differ between the patients, which means that it is impossible to correctly split the data based on this information. With this type of analysis together with consideration of best practice, features that did not contribute to the system with new and unique information were removed. Knowledge from domain expertise also resulted in the removal of useless features. There were 26 features that were removed from Baseline, because they seemed to play a needless role. The features that were removed from Baseline were together called Waste due to their lack of information.

From the database

- Nothing was used in this case

Preprocessing of the waste features

- Identify the content of existing features
- Select which features that do not contribute to information
- Remove these features from baseline and create a set from these

The final feature set from the feature engineering

- Waste (26 features in total)

3.4 Test Engineering

Tests were made with different feature constellations. Since the project was limited in time, not all possible test constellations could be made. Different tests were done with the intention to change the input data to the reviewed system. Features were both added to and deleted from the original Baseline in order to improve the system's ability to make correct predictions. The Baseline worked as a reference, to investigate whether the tests became better or not. The result from Baseline was collected from the test environment based on an in-house version of the predictive AI system. From the tests, AUC values at different prediction days were calculated. A mean AUC value was calculated from these multiple (48) AUC values since they differ over time. This mean AUC value was then compared with the mean AUC value from Baseline.

From the feature construction part of the project, 21 feature sets were created. These sets together contained about 250 features. When a new feature set was created, multiple tests about the set were formed. As all the created feature sets received the same shape in the feature engineering part, they could easily be merged together into a big feature matrix. Different combinations and tests were performed in order to see how the AUC values changed over time, related to Baseline. It was in total 29 tests done during the entire project and the distribution of tests from the different feature groups can be seen in Table 3.6. Some of the feature sets belonged to each other and were therefore always tested together, which was the reason why the number of tests is not so much greater than the number of sets. Each test took approximately two hours before it was completed and because of that a number of tests were generally running after the work day, and ready to be analyzed the day after.

Table 3.6: The test distribution among the feature groups.

Feature group	Number of tests	Test group
Medication	9	M
Dialysis	7	D
Time since healthcare event	2	T
Laboratory results and diagnoses	8	L
Combinations of different groups	3	C

Table 3.7 - 3.11 describe the different tests and which feature sets they consist of. The test names were given from which feature group the feature sets used in the test came from, see test group in Table 3.6. One test included a feature set containing all the new created features in this project and this set is called AllNew. OldMeds, BPOld and OldTimeSince describe the existing medication features, blood pressure feature and the time since healthcare event features from Baseline.

Table 3.7: List over the medication tests.

Test name	Used feature sets
M1	Baseline + MedAdm + MedPre
M2	Baseline - OldMeds + MedAdm + MedPre
M3	Baseline - OldMeds + MedMerged
M4	Baseline - OldMeds + MedAdm
M5	Baseline - OldMeds + MedPre
M6	Baseline - OldMeds + MedMerged + MedMergedRolling
M7	Baseline + MedMerged
M8	Baseline + MedMergedRolling
M9	Baseline + MedAdm + MedPre + MedAdmRolling + MedPreRolling

Table 3.8: List over the dialysis tests.

Test name	Used feature sets
D1	Baseline + BPSitting
D2	Baseline + BPStanding
D3	Baseline + BPStanding + BPSitting
D4	Baseline + BPStanding + BPSitting + BPRolling
D5	Baseline + BPStanding + BPSitting + BPRolling - BPOld
D6	Baseline + BPstanding + BPsitting + BPRolling + WDiff + BFDiff + WRolling + BFRolling
D7	Baseline + WDiff + BFDiff + WRolling + BFRolling

Table 3.9: List over the time since healthcare event tests.

Test name	Used feature sets
T1	Baseline + TimeSinceSaveTop3 + TimeSince
T2	Baseline + TimeSince

Table 3.10: List over the laboratory results and diagnoses tests.

Test name	Used feature sets
L1	Baseline + ICDAll
L2	Baseline + ICDHosp
L3	Baseline + ICDER
L4	Baseline + ICDAll - OldTimeSince
L5	Baseline + ICDAll + ICDAllRolling
L6	Baseline - Waste
L7	Baseline - Waste - OldMeds + MedMerged
L8	Baseline + Age

Table 3.11: List over the combination tests.

Test name	Used feature sets
C1	Baseline - OldMeds + MedMerge + Age
C2	Baseline + AllNew
C3	Baseline + BPStanding + BPSitting + BPRolling + MedMerged + TimeSince

A large part of the project consisted of designing and performing the tests. As previously said, when the new sets of features were created or when old features were removed from Baseline, best practice worked as a guideline. When it came to creating tests the new feature matrices were needed to be incorporated into the existing test environment. It took a lot of time to define the different sets and methods used by the system during the predictions. The number of trees was predefined in the test environment as well as the 48 prediction dates.

3.5 Validation

In this part of the project validations of the results' truthfulness were done in order to understand the outcomes of the tests. The additional aim of this part was to find the strength and weakness of the project.

Another part of validation was to draw conclusions and find connections and reasons why the results were better or worse than Baseline. More tests were created in order to do this. The different aspects which were interesting to study were the influence of each patient category and if the data sets were balanced or not.

3.5.1 Prove Significance

To investigate if a test performed significantly better or worse than the Baseline, Z-test was used. Z-test is a statistical test which builds on the idea of a hypothesis test. This kind of test uses a hypothesis and the thought is to prove whether this hypothesis is probably true or probably not true.

In order to be able to run a Z-test, some criteria are needed to be fulfilled. The sample size must be greater than 30 and this criterion was met as all of the tests contained 48 prediction rounds. The samples should be independent which means that one sample should not affect the outcome of another sample or vice versa. As there was some time period between each day of prediction, it was assumed that the outcome from the 48 predictions did not affect each other. It is also important that the samples which build up the hypothesis are normally distributed. This was assumed by the use of the central limit theorem, which says that the sum of multiple random variables or samples goes towards a normal distribution, although the variables or samples do not have a normal distribution by themselves. The central limit theorem can be used at larger sample collections, which typically contains more than 30 samples, as in this case.

When the needed assumptions were done in order to be able to perform the Z-test the hypothesis was formulated. For each test and time, the difference between the AUC value for the test and the AUC value for the Baseline was calculated. The assumption was that each difference is a sample from an unknown distribution, called X. As said the distribution of X was unknown but the mean μ and variance σ for the distribution X could be estimated as $\hat{\mu}$ and $\hat{\sigma}_n^2$. By using the assumption that the size of X was large enough, $\hat{\mu}$ was calculated as in Equation 3.4.

$$\hat{\mu}_n = \frac{1}{n} \sum_i^n x_i \quad \text{a.s.} \quad \mu \quad \text{as} \quad n \rightarrow \infty \quad (3.4)$$

Where the x_i is a sample of X and n is the number of samples in X. $\hat{\sigma}_n^2$ was calculated by Equation 3.5.

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_i^n (x_i - \hat{\mu}_n)^2 \quad (3.5)$$

Thereafter the confidence interval for the estimation of μ was obtained by using the Z-test. It was decided to use a confidence interval with 95 % probability and therefore the confidence interval of $\hat{\mu}$ could be calculated as in Equation 3.6.

$$\hat{\mu}_n \pm Z_{97.5} \hat{\sigma}_n \quad (3.6)$$

Where $Z_{97.5} = 1.96$ describes the 97.5th percentile of a normal distribution, due to the 95 % confidence interval.

From all these calculations a confidence interval was received for every test performed during the project. If the interval was above zero, the conclusion was drawn that the test was significantly better than Baseline and if the interval was strictly negative the test was considered to be significantly worse than Baseline. If the confidence interval was both positive and negative, i.e., containing zero, conclusions could not be drawn whether the test was better or worse.

3.5.2 Balanced Data Set

The patients within the data can be divided into two groups depending on whether they will be hospitalized or not hospitalized. This is the only two labels that the patients can be classified under. If one of the two labels are more frequent than the other, the dataset can be explained as unbalanced and the classifier learns to recognize this label better since more information about this group is available. Therefore the number of patients in each group was interesting to review. The hypothesis was that the result does not become better if one of the two groups is a majority within the data set.

This was analyzed by splitting the patients into two groups, one group for the hospitalized patients, HP, and one for the not hospitalized patients, NHP. At each training occasion, the system gave information about the labels which could separate the patients into HP and NHP. The methodology, in this case, was to see if the proportion between HP and NHP can be connected to the results.

The correlation between the shape of the AUC values and the differences between the training and testing data were analyzed. This was done by calculating the proportion of HP in the two data sets, i.e., the training and testing set, and thereafter taking the difference of these two values. The result from this was plotted together with the AUC values from Baseline.

To see if balanced data set will change and improve the results, balanced data sets were created. The definition of a balanced data set was in this project to have the same amount of patient in each label category. This was done through a random resampling of the training data set. Due to the randomness, it was not possible to affect the content of the selected training data. The resampling was only done for the training data set, since if the test data set was resampled or balanced it would mean an easier test for the classifier.

In order to balance the training data, the distribution of the two patient groups was identified at each training date. The resampling was done in two ways, in the first way the NHP was reduced to the number of HP, i.e., downsampling. The next way was to increase the number of HP and the method to do this was to duplicate patients from the HP group to the same amount as NHP, i.e., upsampling. Both the up- and downsampling created uniform sized groups at each training day. After the resampling, the test environment worked as before and balanced datasets were tested with Baseline and the feature set containing all, within this project, newly created features, i.e., test C2.

3.5.3 Influence the Results

In this part of the project, tests were done in order to manually influence the results. These tests were done in addition to the other test explained in Section 3.4. The reason for this was to understand how the data should look to be able to do perfectly prediction, but also how the system reacts when only random data is used as input. The first test was in this case made with two completely separated patient groups. This was done by changing all the values for the NHP to zero and keep the HP' original values. The second test was done with only one feature, which gave random values for each patient identification number and date. Except these changes in input data, all the parameters were the same in the test environment.

It was also important to understand how much impact each feature group had on the prediction performance. In order to investigate the influence from each feature group, tests were made where one feature group was removed from Baseline respectively. Tests were also created with only one single feature group used at each time, both for the old and new feature groups. To understand which of the old and new feature groups that performed the best predictions, Z-test was applied between them. In these cases, the tests with the old feature group were used as the reference.

4.1 Identification

In the first part of the identification section, all existing data within the database were investigated. Information about different patient data was summarized and is shown in Table 4.1. The distribution of males and females were almost the same. The interquartile range (IQR) describes the middle 50 % of the population, which for the age means that 50 % of the dialysis patients are between 58 and 79 years old. The reviewed AI system only does predictions for hemodialysis patients qualified for ESCO, but within the database, the primary modality differs among the patients, which also can be seen in Table 4.1.

Table 4.1: Characteristics and distribution of the patients in the database.

Variable	Value
Median age, years (IQR)	69 (58-79)
Sex, number (%)	
Male	3774(53.8)
Female	3176 (45.3)
Unknown	65(0.9)
Median time in dialysis, months (IQR)	57 (25-96)
Primary modality, number (%)	
Hemodialysis	6546(93.5)
CAPD	140(2.0)
APD	26(0.4)
Not mapped	288(4.1)

Table 4.2 shows the different types of healthcare events that the patients do, except their regular dialysis treatments. As can be seen in Table 4.2, the most common visit is the Emergency Department Visit Resulted in Admission which results in hospitalization. These events are unplanned visits and are the only available visits in the database except the planned dialysis treatments.

Table 4.2: A distribution over the different events where the patients in the database seek acute healthcare.

Type of event	Distribution
Emergency Department Visit Resulted in Admission	43.5%
Hospitalization	38.7%
Emergency Department Visit	17.3%
Observation Admission	0.5%

When patients do unplanned visits they normally seek for emergency care. The most common diagnosis that the patients receive during these types of visits can be seen in Table 4.3. The results show that the most frequent problems in addition to ESRD are chest pain and sepsis.

Table 4.3: The five most frequent diagnosis among ESRD patients, sorted in decreased order.

ICD-10 code	Description
R079	Chest pain
A419	Sepsis
E875	Hyperkalemia
R0602	Shortness of breath
I509	Heart failure

The identification part resulted in knowledge about the patient group which was central to this project. This part also contained learning about the reviewed AI system and the training environment as well as insights about which kind of information that was available within the database. The information that was of particular interest in this project contained patient data, dialysis treatments, laboratory results and different kinds of prescribed values.

4.2 Best Practice

The questions about the dialysis work were sent to a contact person at CDC who ensured that they became answered by healthcare staff members with the relevant qualifications. The answers from CDC were summarized and the results from the interview can be seen below. Since the answers were based only on the healthcare staffs' personal expertise the results from the questions cannot be seen as a statistical review. This will, however, give an understanding of the dialysis work and was something that was used during this project.

Dialysis and the daily practice

Dialysis is in some aspects unique in its field of treating patients, the first and most distinct signature is the number of times that the patients visit the clinics each week due to the dialysis treatments. Many different people work at a dialysis center, with different backgrounds and job duties. Around every patient is a primary team with a nephrologist, a social worker, a registered nurse (RN), a dietician and a Care

Manager (CM). The CMs are a special kind of RNs and coordinate the patients' care with all providers that the patients may have. They are also involved in the transitions when the patients are moved in or out of acute care. The RNs are both responsible and accountable for the care during the dialysis treatments.

The personnel within the primary team know the patients a little closer and are present when care plans for the patients are compiled. In addition to these staff members, other RNs and technicians are present as a help at dialysis treatments. The patients usually visit the clinic three times a week to go through dialysis. If the patients' states condition is getting worse, which is interpreted from the patients' symptoms, extra dialysis treatments can be ordered by the nephrologists.

Before a normal dialysis treatment, the technician sets up the dialysis machine according to the orders from the physician. In order to eliminate failures, the technician follows special documentation and then the RN validates the machine as a safety step and completes after that a nursing assessment. During this, the patient weighs him- or herself. When everything is prepared and the patient is placed in the chair, the technician begins the treatment. If a problem arises during the treatment the RN notifies the physician which decides how to proceed.

The patients meet the nephrologists at least once a month but as the patients visit the clinic several times a week due to the dialysis treatments, they meet a lot of other healthcare personnel. The patients can choose anyone from its primary team to talk with and this staff member is there if the patient needs to talk. At the dialysis clinic there are about 10 to 16 patients per each RN, but at days when the staff is unusually short, it can be up to 20 patients per RN. In addition, to this there is one technician per four to five patients.

Comorbidities, tests and routines

When a new patient comes to the clinic a so-called RN assessment is done, where the whole body system is assessed. Most patients suffer from ESRD due to diabetes or hypertension and once the dialysis treatment is started they often have issues with comorbidities like anemia, blood pressure, elevated potassium and fluid overload.

New dialysis patients are noticed by a primary care physician, PCP, who will refer the patient to a nephrologist. The nephrologist determines when the patient have to start the dialysis. Another way of starting dialysis is to show up in the emergency department and there the personnel evaluates if the patient is in need of dialysis. The evaluations are based on the patient's health status, symptoms, laboratory results and also the physician practice preference.

When it comes to hospitalizations non-adherent patients and patients which is near the end of life tend to be admitted more than other. A non-adherent patient is one that is not engaged in the plan of care, even though they have signed it. Typical patients that get more complaints than other are these non-adherent patients, but also diabetic patients which struggle more once on dialysis.

Patients go through monthly tests, where they do laboratory measurements for anemia, bone health, electrolytes in order to see how well the dialyzing works. They also monitor access viability on a regular basis. During a dialysis treatment the blood flow, ultrafiltration rate, dialysis flow rate and vital signs are monitored. In addition to that, tests such as hematocrit (Hct), reticulocyte hemoglobin (Retic HGB) and glucose (Glu) are performed. Before a treatment machine safety tests are completed. Regularly water quality and safety are tested too.

As the staff meets the patient's three times a week they know when something is not right and will then initiate the appropriate actions. Physicians have seen that

there also are other variables except laboratory tests that may affect the general conditions of the patients. These are among other, smoking, single status, food, weather and bad sleep quality. A big piece of health that are not addressed are likewise the social determinants.

Evaluation of patients and the future

There is a learning curve for new nurses and the experienced nurses are typically more comfortable with treating patients with more and severe symptoms. In Ohio, they have a rule that if there only is one nurse at the dialysis unit, it must be a RN with at least one year of nursing experience.

The interviewee thinks that implantable kidneys will play a major role in the future. Also, care coordination is going to become more important for dialysis providers. This in order to see what is happening to the patients outside of dialysis and to improve coordinate care.

4.3 Feature Importance

The feature importance tells which features that have been of high value for the reviewed AI system when the predictions were done. Every feature receives a rank from zero to the number of features that have been used during a prediction. A low value indicates that the feature has high importance for the reviewed AI system and was more important for the prediction than higher values. The feature important rank, which can be seen in Table 4.4, is the mean value of all ranks that a feature acquired during a test, which contained 48 predictions. Every value in the ranking list is individual for the specific test and cannot be compared fairly between different tests.

Table 4.4: Feature importance rank from Baseline.

Rank	Feature name	Mean feature importance
1	TimeSinceHosp0	0
2	TimeSinceHosp1	1
3	TimeSinceHosp2	2.146
4	PostWeight	3.896
5	PreWeight	4.458
6	TimeSinceHosp3	5.458
7	Hct	5.646
8	LastWeight	5.958
9	TreatmentLength	8.250
10	Retic HGB	9.854
11	AverageBloodFlowRate	10.042
12	RunLowBPSystolic	14.375
13	RunLowBPDiastolic	15.042
14	StartSittingBPSystolic	16.229
15	Glu	16.396

Table 4.4 tells which features that have been most important for the system when it was tested with Baseline. The 15 most important features are shown. All features that are in the Baseline set comes from the reviewed AI system.

Table 4.5 shows the 15 most important features from the test when the predictions were made by using Baseline together with all the new features that have been created in this project, test C2. As a result in the comparison between Table 4.4 and Table 4.5 nine features that have been created in this project are placed at top 15 of the most important features. The other six features that remain from Baseline come from the feature groups time since healthcare event and laboratory results and diagnoses.

Table 4.5: Feature importance rank from the test with all created features. A star, *, indicates it is a new feature created in this project.

Rank	Feature name	Mean feature importance
1	Hosp/Week *	0.208
2	TimeSinceHosp0	0.792
3	TimeSinceHosp1	2.042
4	WDiffRoll *	3.412
5	TimeSinceHosp2	4.625
6	BFDiffRoll *	5.583
7	StartSittingBPdiastolicRoll *	5.958
8	EndSittingBPdiastolicRoll *	7.438
9	StartSittingBPsystolicRoll *	7.771
10	DiffDiastoleSittRoll *	10.229
11	EndSittingBPsystolicRoll *	10.271
12	TimeSinceHosp3	10.625
13	Hct	11.771
14	DiffSystoleSittRoll *	12.125
15	Retic HGB	13.604

Features were created in some way within all the four groups that the Baseline was divided into, which is described in the method. To be able to compare the old features from a group with the new features, Figure 4.1-4.4 picture the feature importance rank for all these features. These four figures come from test C2. The blue bars indicate that a feature comes from the old feature sets used in Baseline and the green bars indicate the new features created in this project. From all groups, the 20 highest ranked features were selected to be present within the figures. As there were only three new features in the time since healthcare event group, the three highest ranked features from the old time since healthcare event features were plotted in this case.

Figure 4.1 shows the relation between the old and the new features from the medication group. As can be seen, many of the green bars were placed on a higher rank than the blue ones.

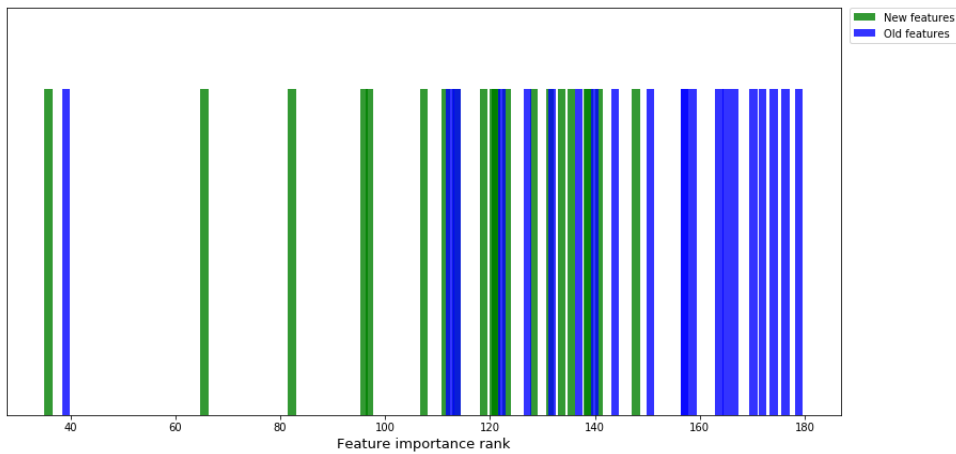


Figure 4.1: Mean feature importance for new and old medication features.

Figure 4.2 shows the feature importance of old and new features from the dialysis group. As can be seen, the majority of the green bars were ranked before the blue bars.

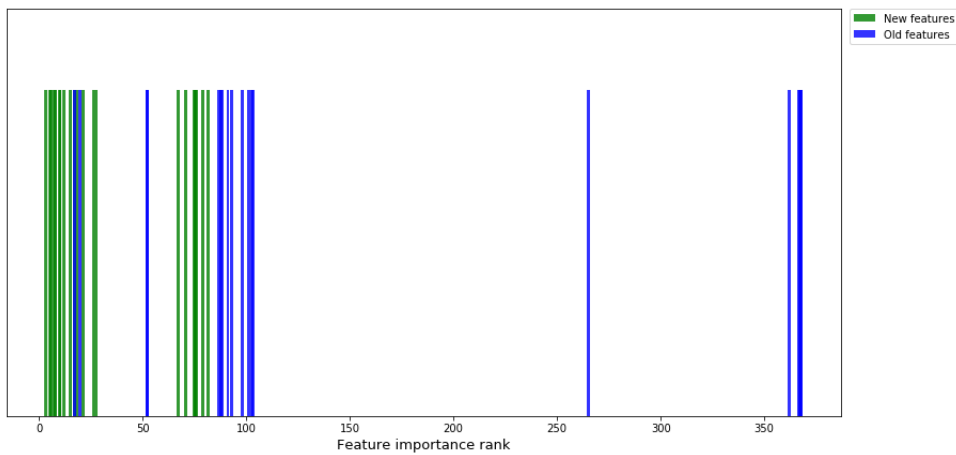


Figure 4.2: Mean feature importance for new and old dialysis features.

The differences in feature rank between old and new features from the time since healthcare event group are shown in Figure 4.3. One of the green bars was placed on the first rank and the other two were ranked after the selected old features, i.e., the blue bars. The feature which was on the first rank describes the hospitalizations per week and patient.

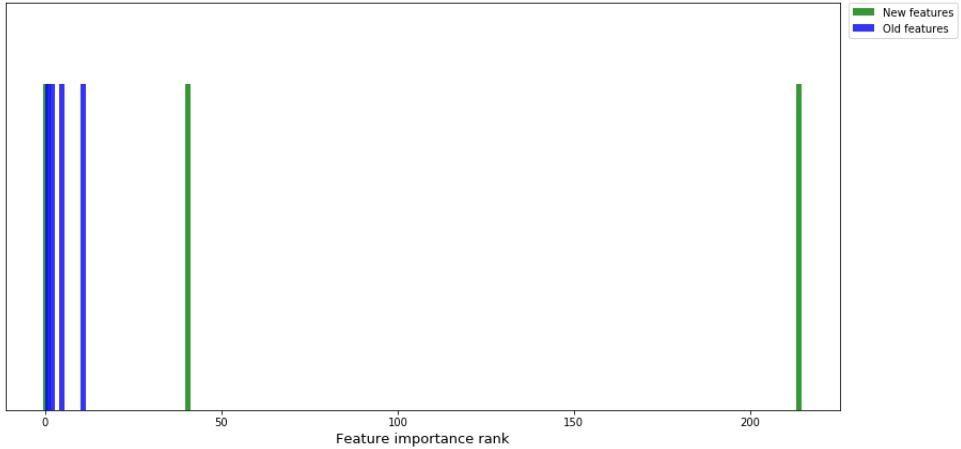


Figure 4.3: Mean feature importance for new and old time since healthcare event features.

The last group that the feature importance rank was investigated on was the laboratory results and diagnoses feature group. As can be seen in Figure 4.4 most of the newly created features were ranked lower than the old features. However there was one green bar which became ranked quite high in the ranking list, and this feature contains information about the patients' ages.

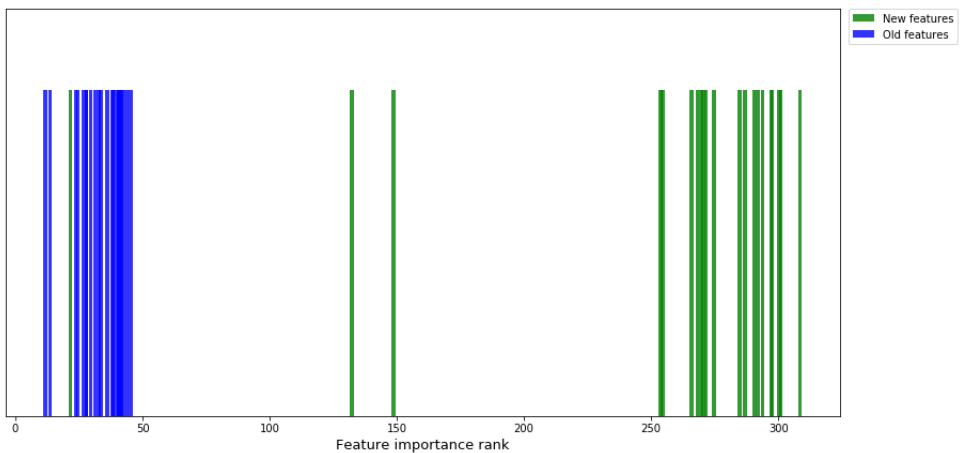


Figure 4.4: Mean feature importance for new and old features from the laboratory tests and diagnoses group.

4.4 Test Results

For each test, a curve which describes the variation of the AUC values over time was received. The AUC values were summarized from the entire test period, which resulted in a mean AUC value for all the 48 predictions performed during a test. A complete summation of all the tests is presented in Appendix B. All the test results with new created features and combinations of them are presented in Figure 4.5. The results are placed under the corresponding test category in the figure, where each category belongs to a specific feature group, see Table 3.6. C stands for combinations, D for Dialysis, L for laboratory results and diagnoses, M for medications and T for time since healthcare event. The red line shows the result from Baseline, which is used as a reference in this project. The mean AUC value from Baseline was 0.727135. As can be seen, many of the tests are placed above the result from Baseline. The mean AUC values were just slightly better on the third decimal.

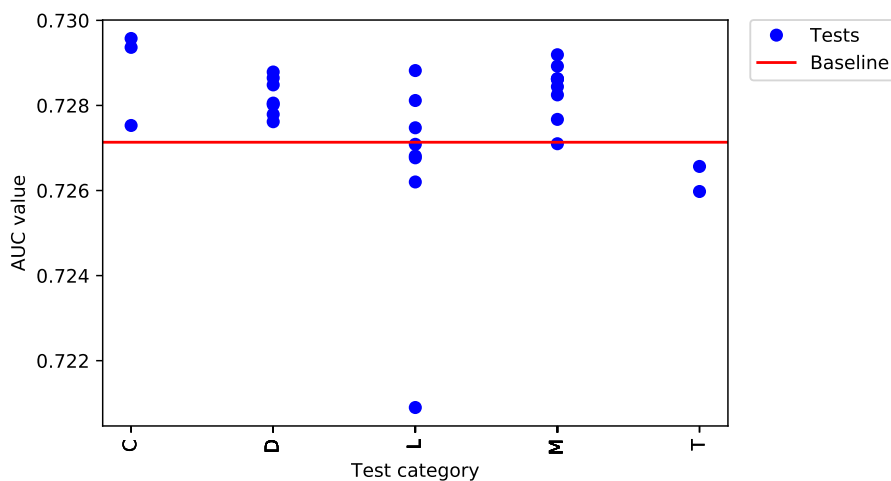


Figure 4.5: Test results from all the tests, grouped after the five test categories. These were the combinations, dialysis, laboratory results and diagnoses, medications and time since healthcare event.

Figure 4.6 shows the variation of the AUC values over time, from the tests Baseline, C2, and L4. All the curves have similar shape even though their AUC values vary in relation to each other. The mean AUC values from the tests in Figure 4.6 are presented in Table 4.6. C2 was the test performed with highest mean AUC value in this project and L4 was the test with lowest mean AUC value.

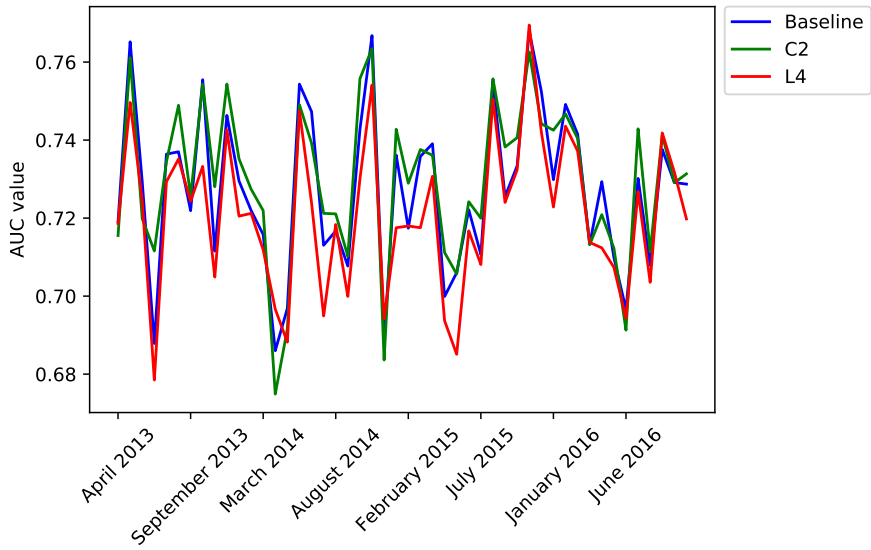


Figure 4.6: AUC values over time for tests with Baseline, C2 and L4.

Table 4.6: This is the mean AUC values from the tests presented in Figure 4.6. The rank of tests describe the mutual order among the test within this project. Test C2 was the best performed test and L4 was the worst.

Test Name	Mean AUC values	Rank of tests
Baseline	0.727135	23 of 29
C2	0.729574	1 of 29
L4	0.720899	29 of 29

4.5 Validation

4.5.1 Prove Significance

To be able to see if the tests were significantly better than Baseline, Z-test was used with a 95 % confidence interval. The results can be seen in Figure 4.7, together with the deviation in AUC values between the tests and Baseline. The tests with green dots became significantly better than Baseline as the confidence interval was strictly positive and the red dot shows a test with strictly negative confidence interval which was, therefore, significantly worse than Baseline. The orange dots represent tests that have unsure intervals and these tests can be considered as neither better or worse than Baseline.

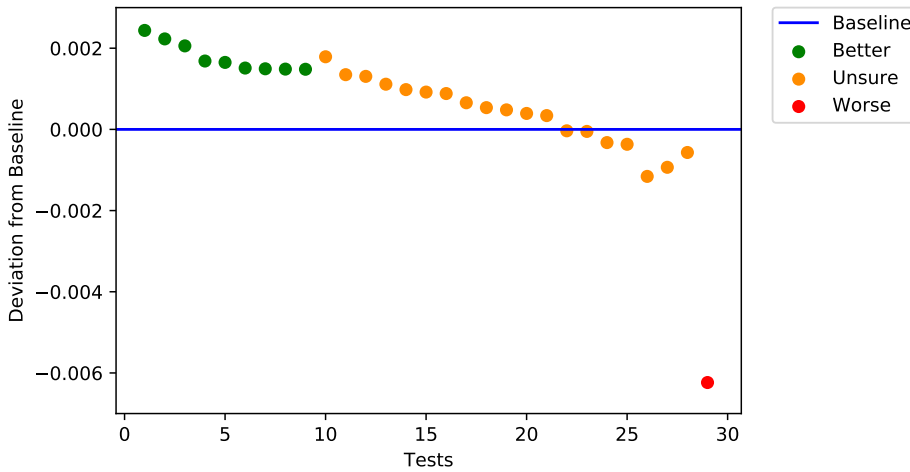


Figure 4.7: Deviations between performed tests and Baseline are shown together with the results from the Z-tests. The colors of the dots indicate if the result was significantly better, worse or unsure.

4.5.2 Distribution of the Patients

The distribution between the different patient groups, HP and NHP, in the testing data are explained in Figure 4.8. It describes the distribution from each day the reviewed system did the predictions. As can be seen in Figure 4.8, there was a relatively similar proportion over time between HP and NHP. The ratio was approximately 1:7 and the data consisted of 551 % more NHP than HP, as can be seen in Table 4.7. The result was that the HP was a minority and the NHP a majority of the patients in the data set that the system was doing the predictions on.

Table 4.7: The mean distribution and standard deviation of testing data between patients from the different groups.

Description	Mean \pm std
Not hospitalized patients, NHP	1016 \pm 77
Hospitalized patients, HP	156 \pm 32

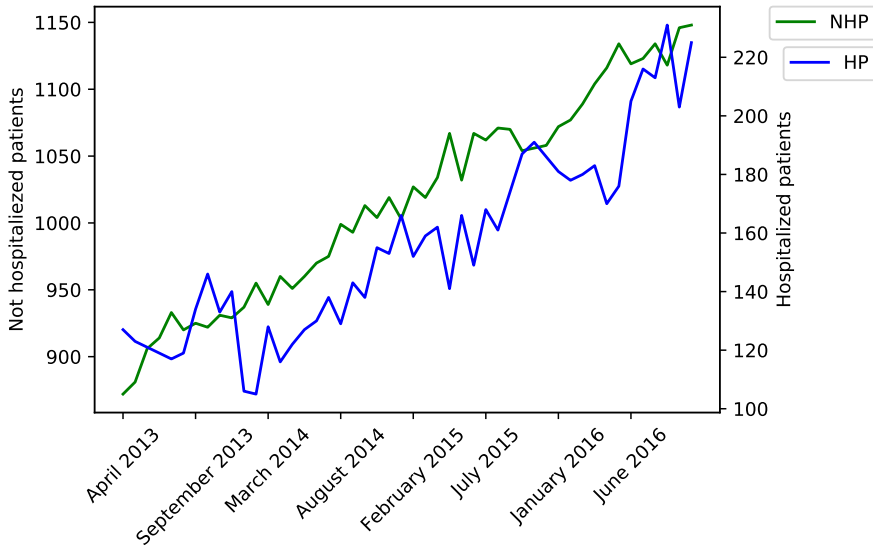


Figure 4.8: The variation between the amount of the two patient groups, NHP and HP, over time in the testing data.

The same analyze was done for the training data, where a similar result was collected. In Figure 4.9 it can be seen that the amount of NHP was always larger than HP, and the ratio was approximately 1:12, see Table 4.8. This resulted in 1068 % more NHP than HP, when the mean distribution where compared within the training data.

Table 4.8: The mean distribution and standard deviation of training data from the different patients groups.

Description	Mean \pm std
Not hospitalized patients, NHP	183281 \pm 54890
Hospitalized patients, HP	15682 \pm 6842

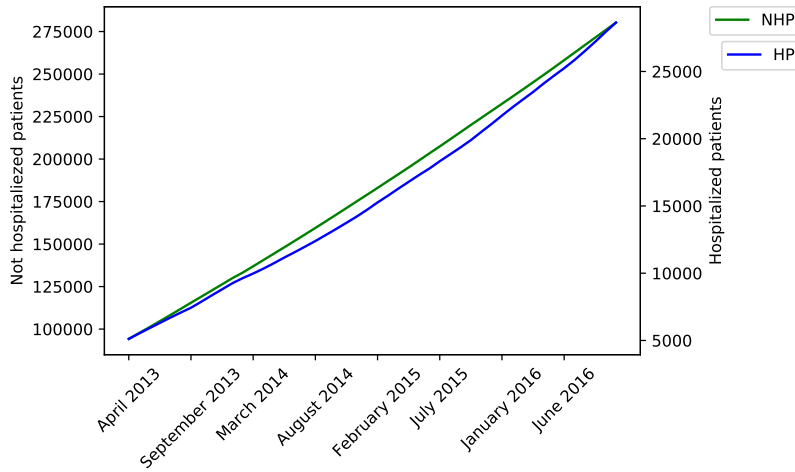


Figure 4.9: The variation between the number of the two patient groups, NHP and HP, over time in the training data.

As can be seen in Figure 4.6, the shape of the AUC values over time looks similar in all the tests. To see if the distribution of HP and NHP could explain the appearance of the AUC values, HPDiff was created. HPDiff is the difference between the proportion of the HP in the testing data and training data over time. This is plotted together with the AUC values over time from Baseline, see Figure 4.10. The variation of the two graphs showed similarities.

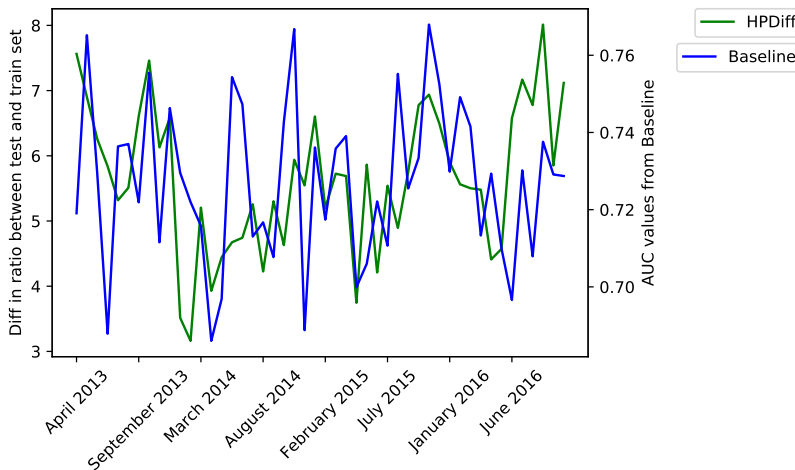


Figure 4.10: HPDiff explains the difference in proportion of HP between testing and training data. The plot includes HPDiff together with the AUC values over time from Baseline.

4.5.3 Balanced Data Set

To see if a balanced dataset could make the reviewed AI algorithm's performance better, tests were created. A balanced dataset means, in this case, that the amount of HP and NHP was the same in every training session. Tests were both made on Baseline and on test C2, where all the newly created features in this project were included. The results are presented in Table 4.9. The table also includes the results from the original datasets. As can be seen, the mean AUC value from Baseline became higher when downsampling was used but lower when upsampling was used. The mean AUC values from C2 became lower both with use of downsampling and upsampling. In Table 4.9 the results from the Z-tests are also included. In both downsampling tests, the results are unsure better than Baseline, due to the Z-test. For the upsampling cases, both tests became worse than Baseline.

Table 4.9: Tests with balanced data sets, mean AUC-values and results from Z-test. Better, unsure and worse describe the significant status with 95% confidence interval.

	Mean AUC	Better	Unsure	Worse
Original				
Baseline	0.727135	-	-	-
C2	0.729574	X		
Downsampling				
Baseline	0.728703		X	
C2	0.726015		X	
Upsampling				
Baseline	0.719467			X
C2	0.721773			X

4.5.4 Influence the Results

The test which is presented in Figure 4.11 is when the two patient groups were separated through modification of the NHP values. The HP keep their original values from Baseline's feature matrix and all the values for NHP were set to zero. In this case, the classifier had 100 % correct predictions. When one random feature was used in another test, the AUC values over time varied around 0.5.

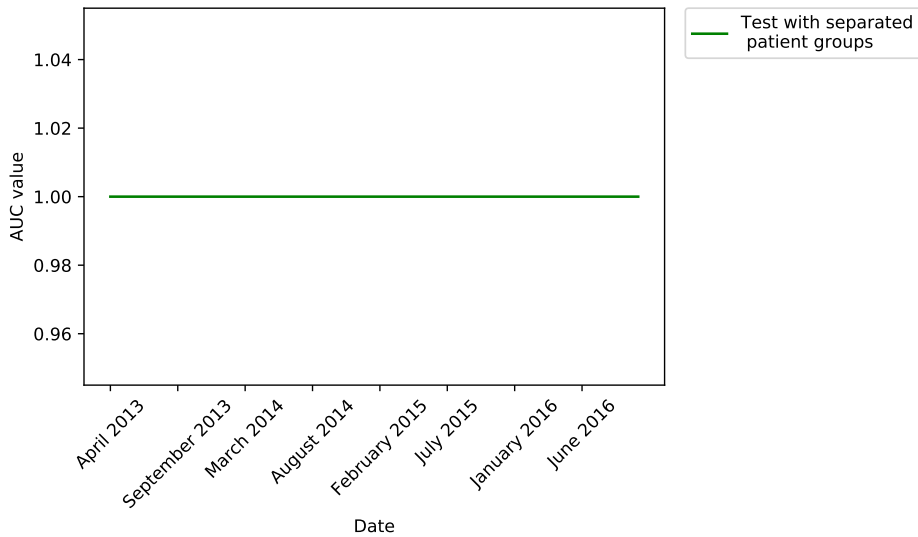


Figure 4.11: Result from the test where the two patient groups were completely separated from each other. The AUC value became constantly 1 which is the value when the classifier has 100 % correct prediction performance.

From the four different feature groups from Baseline, tests were done in order to see which of the groups that play the most important role in the predictions. To do this, one group at each time was removed from the Baseline and the results are presented in Table 4.10. The test where the AUC value became lowest was when the laboratory results and diagnoses group was removed. This was also the test with least features left.

Table 4.10: Tests where one of the feature group is removed from Baseline.

Removed group	Mean AUC	Deviation from Baseline	Number of features in test
Medication	0.726171	-0.000964	251
Dialysis	0.722028	-0.005107	282
Time since healthcare event	0.720215	-0.00692	206
Laboratory results and diagnoses	0.718775	-0.00836	179

When tests were performed with one feature group each, the mean AUC values became lower than it has been before, see Table 4.11. Tests were made both with the old and new feature groups. As can be seen in 4.11 the new medication and dialysis features became significantly better than the old medication and dialysis features. The new time since healthcare event and laboratory results and diagnoses features received significantly worse results in comparison with the old features from these

groups. The Z-test was calculated with the new features in relation to the old ones from each group respectively.

Table 4.11: Tests related to single groups were created, both with old and new sets in each feature group. Deviation is the difference between the mean AUC values from each set.

Feature group	Old set mean AUC	New set mean AUC	Deviation New and Old
Medication	0.650644	0.665929	0.015285
Dialysis	0.646770	0.65666	0.00989
Time since healthcare event	0.655130	0.561418	-0.093712
Laboratory results and diagnoses	0.702424	0.543892	-0.158532

The world's population lives longer and suffers from chronic diseases to a larger extent than ever before. This increases the workload on the hospitals and also the care costs. To be able to continue delivering high performed care, complements, e.g., predictive AI systems, to the common nurse assessments are of great importance. The purpose of the systems is not to replace the existing healthcare staff, they should just work as complements when taking decisions and prioritizing the patients with the most need. The meaning of the systems is to find complex connections that a human eye is unable to detect. In this project, the aim was to make an in-house version of an AI system even better at detecting patients with the most need of care. A lot of time during this project has been spent on discussing whether the newly constructed features made the system better or not.

As can be seen in the test results according to the mean AUC values, the values often just differ on the third decimal. A classification problem that is hard to solve is reasonably also harder to make improvements for better predictions on. There could be lots of more unmeasurable parameters that would make sense for a hospitalization, and if the system does not have access to these parameters, it is difficult for the system to classify correctly.

5.1 The New Features

One of the main goals of this project was to create new features that hopefully could improve the prediction performance of the reviewed AI system. To compare the new features with the existing ones, the feature importance method was used. As can be seen when comparing Table 4.4 and 4.5, many of the newly created features were placed at top 15 on the ranking list which meant that the old features declined on the list of feature importance. From this, the conclusion was drawn that the reviewed AI system preferred to use some of the new features when classifying the patients into the two groups, HP and NHP. This measurement does not present the whole picture of the feature importance, e.g., the interaction between several features. Even though the random forest classifier is good to find these interactions, we were not able to closer investigate this during the project. From the results it is, however, possible to see that the new features affected the feature importance.

As can be seen in Figures 4.1-4.4, many of the new features were placed before the old, existing features from Baseline. It was mainly the new features from the medication group and the dialysis group that receive higher feature importance. One of the newly created features from the time since healthcare event group was always placed first when it was used in the tests. The new features from the laboratory

results and diagnoses group performed worse than the old features, except the feature related to patients' ages.

Possible reasons why the new features were placed quite high in the feature importance rank could be due to that they brought new information to the reviewed AI system. It can also be because they contained values that made it easier to distinguish between the HP and the NHP. We also believe that the use of best practice was an important reason for the results.

Feature importance could not prove all possible interactions between different features, but by adding and removing features we could, as an example, see that the time since healthcare event features affected each other. When adding the feature set TimeSince together with Baseline two of the features were placed on rank 52 and 159. In another test, only the top three time since healthcare event were kept, which resulted in rank 8 and 112 for the same features as above. Even though the second test was done with fewer features the rank cannot be explained only by the loss of features. This implies that the new features contained similar information as the old ones and are therefore ranked lower when the AI system can use all these features together. The change in rank showed some sort of interaction between the new time since healthcare event features and the old ones. The hospitalizations per week feature in this set was always on the first place of the ranking list, regardless the other features used in the test. This was probably due to the fact that this feature contained new important information that the system has not had before.

When looking at the new features that get a low rank, two possible reasons for this can be that these new features did not contribute with new information or that the new information was too unspecified, unspecified in the sense that the information was too general and could not be used to distinguish between the two patient groups. For example, the features in the diagnoses group only used the first character of the ICD-10 code in the feature construction part. This probably caused a too broad grouping where both severe and more harmless complaints ended up together in the same group. Due to this group division, deeper information about the different diseases and complaints became lost.

In addition to the grouping of the diagnoses, it was really difficult to preprocess the ICD-10 codes into realistic features based on best practice. First of all, it was difficult to understand how the information in the database should be interpreted and how it was registered by the clinicians. Second, it was impossible to know how long the patients suffered from each disease and we could therefore not know in which time interval the information should be applied. Because of this, it was understandable that this feature set did not give a desirable prediction performance due to the unsureness in the feature construction.

The diagnoses used in the new features were also diagnoses established during emergency care, i.e., the three first visits that are described in Table 4.2, which do not describe the whole picture of the diagnoses' history for a patient. It would be interesting to survey all the diagnoses a patient suffers from during a specific week, including both hospitalization diagnoses and other diseases in addition to ESRD. To be able to map this out, many different tables are needed to be completed with information, and that command high quality of the data in the database, both the documentation of the patients' clinical pictures, but also that the information is stored in a correct way, based on the dates and the diagnoses.

The existing features in Baseline related to diagnoses only covered a few ICD-10 codes, and the new features were an approach to give the system more information about all patients health problems. Our approach was not good enough, and one

reason for the poor result can be that it was difficult to know for how long a patient has suffered from a disease. By just looking at one patient who has several similar diagnoses during a period of time, it was impossible to translate the information correctly for all cases. This kind of information can be interpreted as a diagnosis that disappears and comes back or as a long-drawn disease which leads to multiple hospitalizations. Anyway, we still think that information about diagnoses should be informative for the system if it tells the truth about all diagnoses a patient has or has had.

When trying to improve the features from the medication group, we tried to put in as much information about the patients' medications as possible. The new medication groups were, therefore, a good approach, to have a realistic amount of features and not using all the medicines as individual features. When comparing tests with and without the newly created medication features, the best result was received when the new medication features were used without the old medication features. This means that the new medication features can both replace old features at the same time as they contribute to new information which increased the prediction performance. This was also confirmed when tests were made with old and new medication features respectively. In these tests it could clearly be seen that the new medication features performed significantly better on their own in comparison with the old features in this group, see Table 4.11.

In order to see which of the old feature groups that had most predictive variables for the reviewed AI system, four tests were done with one feature group each, Table 4.11. The results became worse than both Baseline and all of the performed tests, which indicates that there were some important variables in each group that could contribute to the prediction performance. The opposite was also done, and in this case tests with three feature groups at each time were created, see Table 3.10. From the results, we draw the conclusion that not all of the old feature groups are needed in order to perform quite as good result as Baseline.

The number of features in each test were studied in order to see if there was a correlation between the number of features and the prediction result. Except that the best test, C2, contained the largest amount of features, a correlation could not be found. A larger number of features enables a greater amount of information within a feature set, but if the information is of bad quality for the prediction performance, it does not matter how many features that are used.

5.2 Z-test

As can be seen in the test results according to the AUC values in Figure 4.6, the results do not differ much from the test with Baseline. Many of the tests, however, became slightly better when comparing their mean AUC values and the question was if these improvements were significantly better or not. Only by studying the mean AUC values, it was impossible to prove significance as the differences in the mean AUC values were in the order of a few thousandths. Therefore, the Z-test was used for all tests. The result from this was that nine tests were significantly better, one test was significantly worse and the rest had unsure results from the Z-test. Unsure result means that it was impossible to determine whether a test was better than Baseline or not. The tests that performed better than Baseline, contained new feature sets with information about medication, blood pressure, and age. This confirmed the results from the feature importance.

5.3 The Variation of the AUC Values

When looking at the shape of the curves and the variation of the AUC values over time, in Figure 4.6, it can be seen that the same shape originates for all of the tests. Therefore the conclusion was drawn that this has nothing to do with the work made in this project. Speculations have been made in order to explain this behavior. As a suggestion, we think that this was because of the differences among the two patient groups as well as differences between the training and testing data. The differences can, for example, be due to the number of patients in each group, the variation of diagnoses and external variables that are impossible to measure or at least not available within this project.

In this project we only studied the number of patients in each group, both for the training and testing data, the rest can be seen as future work. Figure 4.9 shows the relationship between HP and NHP overtime for the training data and these curves raise over time. The same pattern can be seen for the number of patients in the testing data, see Figure 4.8. In both cases, the relation between HP and NHP was almost the same at each day of prediction even though they differed much in size. The reason for the increase in training data was that the amount of patients was growing within the database and all the stored information was used at every training session. The increase of patients in the testing data was due to the number of active patients connected to the reviewed AI system.

The outcome from the predictions is dependent on both the training and testing data. If the system is trained on one type or variation of patients, it is difficult for the system to solve the classification problem if the testing data consists of a completely different patient group. In Figure 4.7 it can be seen that the two curves, the AUC values and the differences in the proportion of HP between training data and testing data, follow each other fairly good. The differences in size between the training and testing data could therefore in some way explain the variation of AUC values between the prediction days. However, there are still some abnormalities in the variation of AUC values that cannot be explained by this comparison. What causes this is something outside this project but nevertheless important to understand.

To use the same number of patients from the two patient groups during the training sessions, the data set was balanced both by downsampling and upsampling. The results from these tests did not show any improvements on the prediction performance when validating with Z-test. The downsampling test for Baseline got higher mean AUC value but the Z-test resulted in unsure improvement. It was also an unsure improvement for the test where C2 was performed with downsampling. Information from the NHP group was lost when the training data was downsampled and the amount of training objects was reduced. This was probably the reason why these tests did not improve. Both the tests done with upsampling became significantly worse than Baseline, determined by the Z-test. The upsampling tests were made by duplicating the existing information of the HP which did not help the reviewed AI system to increase the prediction performance. It would be interesting to create balanced dataset with the addition of totally new information from other hospitalized patients.

The AUC value rises particularly in the beginning of each test period in general, see Figure 4.6, and this can be connected to the small set of information and the small number of HP since the information has not been collected for long. As the AUC values differ much over time there must be other reasons that we did not completely find in this project. According to this the attempts to improve the

prediction performance for all 48 occasions were maybe too complex for this project. One approach to solve this problem could be to only analyze one specific session and trying to understand and improve the predictions on this specific day. On the other hand, that would not be a picture of the whole truth since the classifier should be able to handle different types of input data, because that is the reality for an AI system used within the healthcare system.

When looking at the mean AUC values from all of the tests, they did not differ much from each other. This seemed suspicious and was the reason why the validation part contained tests with consciously better and worse sets of input data. In Figure 4.11 it can be seen that it was possible to create two patient groups which can be completely separated by the reviewed AI system. This indicates that the system has difficulties in separating HP from NHP. The reason for that was that the feature values, both for the new and old features, were too similar for the two groups. When just having a random feature as input data, the mean AUC value was around 0.5, which is the value that the ROC curve gives when the classifier is using random input data.

5.4 The Use of Historical Information

The created features according to rolling mean values were implemented since the reviewed algorithm did not use much of historical data about previous feature values, for example, the previous week values compared with this week's values. When looking at feature importance rank, the rolling mean features and the old time since healthcare event features have been placed high, and we interpret that this kind of information is important to the system. It would be really interesting to use even more time series analysis to include patients' earlier values, something we think are of high importance to be able to find deviations in a patient's health. The complexity is still there and obvious, one change in a value can have a big effect on one patient, but no effect at all on another patient. This is because there are so many parameters that matter in a human's health, not least the physical parameters which are hard to measure and something for the system's future.

5.5 The Use of Best Practice

Another version of the reviewed AI system has been used within the healthcare system for over a year and has been proved to have a positive effect on the treatment of patients with ESRD. According to the nurses which use that system in their daily work, the AI system is able to find some patients that will be hospitalized, but as always there is room for improvements. Besides the fact that a better system, which makes reliable predictions is desirable, the output from the system needs to be better too. This is something that is under development, but we think that better features are one way to create an output that is easier to interpret.

Today the output from the reviewed system is rather vague, it is a received AI score between zero and one for each patient. As it is not possible to do an evaluation by just looking at the scores, the patients' scores need to be compared, in order to see which patient that is at greatest risk of becoming hospitalized within 30 days. Even though the AI scores give a guidance on where to put the resources, the reviewed system would be more useful if it could provide more help to the clinical staff. To further improve the usefulness of the reviewed system, the output needs to contain

more information about why a specific patient receives a higher AI score than another patient. In addition to this, the output should also be able to work as guidelines that recommend the interventions that are needed to be done to prevent or prepare the upcoming hospitalization.

It is shown that best practice is a useful way of improving input data to an AI system and even though the prediction performance became just slightly better in this project, we think that the new features can be one way of improving the reviewed system's output. If the features are not preprocessed by using best practice, it can be difficult to show the most important features for the medical staff because these features will probably mean nothing to the personnel. If the features are created with best practice in mind, the nurses will be able to use the features with a more straight forward approach as this kind of features are created from how the nurse assessments are done. Of course, the most important thing is that the features give predictions that are both reliable and accurate. In other words, an AI system's output is more or less useless unless the predictions are telling the truth.

The problem can in advantage be divided into specific cases, in order to improve the predictions for specific groups. The reviewed algorithm looks at the same variables for all patients. ESRD patients often have complex medical histories with one or several comorbidities. Due to that, different groups of patients have different variables that could be important to use for predictions. If it was possible to find important variables for different subgroups of patients that for examples have both ESRD and diabetes or ESRD and heart disease, more specific and individual predictions can hopefully be made.

5.6 Test Structure

All parts of the feature construction and feature selection have been controlled by us, which means that all the feature sets were selected by hand and with knowledge from domain expertise. The whole project was based on the results from Baseline and all tests were different versions of that feature set. The big disadvantage with doing feature constructions by hand is the possible risk to not detect important connections that the system could be trained on, and that information we do not render as important could be useful for the system. Another disadvantage is that we were not sure if we have created the best possible test. We were also aware of that the number of created features were not enough or their content was not sufficient in order to obtain larger improvements.

Many results from patients' measurements were collected several times during a week. When these were connected to week intervals, information was risked to be lost during the feature engineering steps. The way we often handled information occurring more than one time per week, was to take a mean value of all the existing values. Because of this, we lost information especially about quick changes within a week. The new values were estimations of several values. A possible disadvantage with this was that we miss values or changes that could indicate hospitalizations.

5.7 Limitations and Error Sources

As all projects like this, limitations are difficult to completely avoid. The important thing is to identify these possible error sources and be aware of them. The largest factor that has limited the work and in the end maybe also the results was that data

was missing in some cases. Due to this, some features were not possible to create or in other cases we needed to make assumptions to fill in missing data.

The reason for the amount of missing data is difficult to tell with great certainty. The data is collected from different hospitals, where the routines of how to fill in the medical record may differ between each unit. In addition to this, the personnel is all human beings and it is therefore impossible to avoid human errors. During the work with the information within the database, several patients with slightly contradictory case history have been found. Presumably, the healthcare staff make mistakes and instead of changing the medical record, they fill in a new event without removing the incorrect one. It is impossible to find all of these errors and once they are found it is difficult to know what is correct and what is not.

Another limitation of our work, that we have found, is that the different hospitals' EHRs do not communicate flawlessly. The best example of this is how the hospitals report hospitalization events. Some hospitals are very good at distinguishing between emergency department visits and hospitalization admissions while other hospitals merge these events together under the category hospitalization admissions. This made it hard for us to exactly know which kind of hospital events that the patients do, which in the end affect the creation of new features.

Due to the limitations that were identified during this project, we had to do assumptions. As described above, we have tried to think of best practice when data was missing. The same has been applied when incorrect information was suspected to be written within the medical records, i.e., when information was contradictory. It was hard to estimate the impacts of these assumptions and not least, but still most important, which assumptions that brought the results closest to the truth.

5.8 Future Work

The need of an AI system is, according to what we have seen during this project, undoubtedly huge. The upcoming challenge is however to create a system which is able to do correct predictions for all different kind of patients. The attempts made during this project, to find a system like that, have not given the most desired outcome. With this in mind, we think that only new features are not enough to improve the existing system. The focus of this project was to incorporate best practice into the reviewed system's structure. Since this did not succeed in convinced results it would be interesting to see if another approach could improve the possibilities for better prediction performance even more. Larger system changes are probably needed in order to reach this. It would also, among others, be interesting to investigate what happens if the system becomes rewritten with another structure, which especially uses more information that goes back in time. This was however not the purpose of this project since the primary aim was to change the input data after best practice used within the healthcare system.

Work needs to be done in order to completely understand why there are some days where the system receives much smaller AUC values. It is possible that these changes in AUC values are depending on the quality of the data or the number of patients. Another possible reason for the fluctuations of the AUC values is that there are patients whose health status are more difficult to predict. If that is the case, the next step would be to identify these patients and investigate which kind of features that can distinguish them from each other.

Due to the fact that assumptions had to be made when data was missing or

unreliable, it would be interesting to investigate if it is possible to create better data with higher quality. The key to do this is maybe to change the way the data is gathered and saved. As always when it comes to classification problems, a large amount of data are needed. Therefore one step forward in the direction of a better classifier is to collect more data from different sources. One impression from this project is that this specific reviewed AI system needs more and new information to become better. Especially new historical information about patients would be interesting to use as input in order to find different value variations which can indicate on degenerated health. Time series analysis could be a possible improvement to capture the historical changes over time. Finding more data with high quality, together with identification of the patients that are more difficult to predict correct, will hopefully contribute to a better system.

Since there are a lot of reasons to why patients get hospitalized, not just deviated values, it would be beneficial to include other parameters in the system also, such as quality of life data. If it is possible to find more important information where the system can find connections to the patients' health, the system gets a more reliable picture of the situation, and can then hopefully perform even better.

Another thing to do in the future is to group the patients into subgroups dependent on their diseases in addition to ESRD. This should be done to see if there are specific variables that are important and interesting for each group, and something that then needs to be used as input, i.e., the set of features would probably differ between the different subgroups of patients.

To make the reviewed AI system popular among clinics and clinicians, the way the information is presented needs to be more hands-on and clear on the dashboard. One thought is to tell the nurses which parameters the system reacted on and which values it based the decisions on. The removal of the features in the set called Waste was one approach to create a system that only uses intelligible information. If the input data is close to best practice, as we try to imitate in this project, the output data would probably be easier for the users to understand.

5.9 Ethical Considerations

There is a lot to take in consideration when it comes to ethical aspects as this project involved a system that uses both medical and personal information from patients. The hope is that this project will lead to a better healthcare system and one important part in order to do this is to ensure the patients' privacy. There is an agreement with permission to store and analyse the information about the patients within the database. This provided that the usage is in compliance with the guidelines in HIPAA.

Within the database, a lot of information is stored and this is not only medical information. Personal information such as home address, marital status, ethnicity, etc, is also available through the database. This means that a person who accesses the database can know very much about each patient, which sets high security requirements on the system. It is unacceptable that this kind of information falls into wrong hands. The personal information, except the medical history, is not used within the system as it is built today. The question arises if it is defensible to store this amount of unused information. Is there a reason to have all this information when it is not used, or should this information be implemented within the system? Maybe, it is not moral or ethically correct to use personal information to make

predictions, but if it can be proven that the personal information has high predictive value, is it then acceptable to collect and use this kind of information?

When the security within the database is ensured, other aspects are important to discuss. A limitation of this project was missing, or incorrect, information within the database. The system uses this information for the predictions and if the information is incorrect for some reason, the predictions are likely to be wrong. Who is responsible for the information and who takes the consequences if something unexpected happens? The AI system is developed as a decision support, and should also be used as that. What happens if the system's prediction is wrong? If the clinicians completely trust on the system, a wrong prediction can result in terrible consequences. For example, patients can be treated incorrect or patients that are in great need of care may be missed by the system. If these cases occur, who take the responsibility? These are some aspects that one has to be aware of when using or working with systems like this.

With a more developed healthcare system, with advanced technology, the personal contact is at risk of being reduced. Therefore it is important for the users to understand that this AI system only is a resource when making decisions, and the evaluation can not only be based on the predictions.

Author Contributions

We, Elin Edvinsson and Anna Goos, have both contributed equally to the work of this master's thesis. The design and analysis of the tests and the production of the report were sometimes made parallel and in these cases one of us proofread the written code, text or the received results made by the other, in order to ensure good quality of the work.

Conclusions

In this project, about 250 new features were created in order to be used as input data in a predictive AI system. The feature extraction was done with the use of best practice from domain expertise within dialysis care. A test environment was used in order to test the new features. The feature importance method and Z-test were used to evaluate the results. The goal was to improve the features within the system, and even though the prediction performance became just slightly better, many of the new features were placed at a higher feature importance rank than the old ones. The conclusion was drawn that major system changes are needed to be done in order to improve the system from where it is today. Since this was outside the scope of this project, it will be saved as a challenge for the future.

Bibliography

- [1] Abbasi. M, Chertow. M.G, Hall. N.Y (2009). *End-stage renal disease* Clinical Evidence [Online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3217820/pdf/2010-2002.pdf> [Retrieved 2 March 2018]
- [2] Academy of professional coders (AAPC) (n.d.). *What is medical coding?* [Online] Available at: <https://www.aapc.com/medical-coding/medical-coding.aspx> [Retrieved 2 March 2018]
- [3] American Medical Association. (n.d.). *ICD-10 Overview* [Online] Available at: <https://www.ama-assn.org/practice-management/reporting-continued-issues-icd-10-claims> [Retrieved 19 March 2018]
- [4] Breiman. L. (2001). *Random Forest* Machine Learning. 45: 5-32
- [5] Centers for dialysis care, CDC. (n.d.). Northeast Ohio renal alliance LLC. Available at: <http://www.cdcare.org/news/cdc-access-care/> [Retrieved 2 February 2018]
- [6] Centers for Medicare & Medicaid Services. (n.d.). *Comprehensive ESRD care model* [Online] Available at: <https://innovation.cms.gov/initiatives/comprehensive-esrd-care/> [Retrieved 25 February 2018]
- [7] Centers for Medicare & Medicaid Services. (n.d.). *The comprehensive ESRD care initiative* [Online] Available at: <https://innovation.cms.gov/Files/slides/Comprehensive-ESRD-Care-ODF.pdf> [Retrieved 10 March 2018]
- [8] Couchoud. C, Moranne. O, Frimat. K, Labeuw. M, Allot. A, Stengel. B (2007). *Associations between comorbidities, treatment choice and outcome in the elderly with end-stage renal disease* Nephrology dialysis transplantation. 22(11), 3246-3254
- [9] Coyne. D.W (2011). *CKD Medscape CME Expert Column Series: Issue 3 - Management of Chronic Kidney Disease Comorbidities* Medscape [Online] Available at: <https://www.medscape.org/viewarticle/736181> [Retrieved 16 April 2018]
- [10] Dialysis Clinic, Inc: *A better way to coordinate care.* (n.d.). [Online] Available at: <http://www.dciinc.org/escos/> [Retrieved 8 March 2018]
- [11] Enders, C.K (2010). *Applied Missing Data Analysis.* New York: The Guilford Press

- [12] Gorm Analysis (2014). *Magic Behind Constructing a Decision Tree* [Online] Available at: <https://gormanalysis.com/magic-behind-constructing-a-decision-tree/> [Retrieved 23 May 2018]
- [13] Guyon. I, Gunn. S, Nikravesh. M, Zadeh. L.A (2006). *Feature Extraction: Foundations and Applications*. The Netherlands: Springer
- [14] Hall. K.R, Luciano. A, Pieper. C, Colón-Emeric. S.C (2018). *Association of kidney disease quality of life (KDQOL-36) with mortality and hospitalization in older adults receiving hemodialysis* BMC Nephrology. 19(11). 1-9
- [15] Hand. D.J (2009). *Measuring classifier performance: a coherent alternative to the area under the ROC curve*. Mach Learn. 77: 103-123
- [16] Hanley. J.A, McNeil. B.J (1982). *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*. Radiology. 143: 29-36
- [17] Health Network Solutions (HNS) (n.d.). *Anatomy of ICD-10 codes* Available at: <http://www.healthnetworksolutions.net/index.php/understanding-the-icd-10-code-structure> [Retrieved 2 March 2018]
- [18] Honeycutt. A.A, Segel. E.J, Zhuo. X, Hoerger. J.T, Imai. K, Williams. D (2013). *Medical Costs of CKD in the Medicare Population*. Journal of the American society of nephrology. 24: 1478-1483
- [19] Joyce. T.A, Iacoviello. M.J, Nag. S, Sajjan. S, Jilinskaia. E, Throop. D, Pedan. A, Ollendorf. A.D, Alexander. M.C (2004). *End stage renal disease-associated managed care costs among patients with and without diabetes*. Diabetes Care. 27(12), 2829-2836
- [20] Kaushik. S (2016). *Introduction to Feature Selection methods with an example (or how to select the right variables?)* Analytics Vidhya.[Online] Available at:<https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/> [Retrieved 5 May 2018]
- [21] Keith. S.D, Nichols. A.G, Gullion. M.C (2004). *Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization* Jama internal medicine. 164(6), 659-663
- [22] *Learn more about medical coding* (n.d.). [Online] Available at: <http://www.medicalbillingandcoding.org/learn-more-about-coding/> [Retrieved 2 March 2018]
- [23] Lin.Y-T, Wu. P.H, Kuo. M-C, Lin. M-Y, Lee. T-C, Chie. Y-W, Hwang. S-J, Chen. H.C (2013). *High cost and low survival rate in high comorbidity incident elderly hemodialysis patients* PLoS ONE 8(9): 1-8
- [24] Long. B, Koyfman. A, Lee. M.C (2017). *Emergency medicine evaluation and management of the end stage renal disease patient* American Journal of Emergency Medicine. 35, 1946-1955
- [25] Lovansik. P.B, Zhang. R, Hockenberry. M.J. (2016). *Emergency Department Use and Hospital Admissions among patients with end-stage renal disease in the united states*. Jama internal medicine. 176(10), 1563-1565

- [26] LumenCandela (n.d.). *Physiology of the kidneys* [Online] Available at: <https://courses.lumenlearning.com/boundless-ap/chapter/physiology-of-the-kidneys/> [Retrieved 2 March 2018]
- [27] MEDCALC (2018). *ROC curve analysis* [Online] Available at: <https://www.medcalc.org/manual/roc-curves.php> [Retrieved 11 May 2018]
- [28] MedicineNet.com (n.d.) *Creatinine (Low, High, Blood Test Results Explained)* [Online] Available at: https://www.medicinenet.com/creatinine_blood_test/article.htm#what_is_creatinine [Retrieved 28 May 2018]
- [29] MedlinePlus (2018). *Glomerular filtration rate* [Online] Available at: <https://medlineplus.gov/ency/article/007305.htm> [Retrieved 2 March 2018]
- [30] National Kidney Foundation (2017). *About Chronic Kidney Disease* [Online] Available at: <https://www.kidney.org/atoz/content/about-chronic-kidney-disease> [Retrieved 12 March 2018]
- [31] National Kidney Foundation (2015). *Dialysis* [Online] Available at: <https://www.kidney.org/atoz/content/dialysisinfo> [Retrieved 2 April 2018]
- [32] National Kidney Foundation (2018). *Glomerular Filtration Rate (GFR)* [Online] Available at: <https://www.kidney.org/atoz/content/gfr> [Retrieved 21 February 2018]
- [33] National Kidney Foundation (2017). *How your kidneys work* [Online] Available at: <https://www.kidney.org/kidneydisease/howkidneyswrk> [Retrieved 17 February 2018]
- [34] Navaneethan. D.S, Jolly. E.S, Schold. D.J, Arrigain. A, Saupe. W, Sharp. J, Lyons, J, Simon. F.J, Schreiber. J.M, Jain. A, Nally. V.J (2011). *Development and validation of an electronic health record-based chronic kidney disease registry*. Clinical journal of the American Society of Nephrology. 6, 40-49.
- [35] NIDDK (2008). *Hemodialysis. National institute of diabetes and digestive and kidney diseases* [Online] Available at: <https://www.niddk.nih.gov/health-information/kidney-disease/kidney-failure/hemodialysis> [Retrieved 21 March 2018]
- [36] Office for Civil Right (OCR) (2017). *HIPAA for professionals*. [Online] Available at: <https://www.hhs.gov/hipaa/for-professionals/index.html> [Retrieved 1 February 2018]
- [37] Persaud. N (2016). *ESRD patients visit emergency departments more frequently* Renal and Urology news. [Online] Available at: <https://www.renalandurologynews.com/secondary-hyperparathyroidism/esrd-patients-visit-emergency-departments-more-frequently/article/518928/> [Retrieved 21 February 2018]
- [38] Random Forests, Leo Breiman and Adele Cutler (n.d.). *Gini importance* [Online] Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#giniimp [Retrieved 23 May 2018]
- [39] Robert. W. Schrier. M.D, Wei Wang. M.D (2004). *Acute Renal Failure and Sepsis*. The New England Journal of medicine. 351, 159-169

-
- [40] scikit-learn (n.d.). 3.2.4.3.1. *sklearn.ensemble.RandomForestClassifier* [Online] Available at: <http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> [Retrieved 16 May 2018]
- [41] Sondhi. P (2009). *Feature Construction Methods: A Survey* [Online] Available at: <http://sifaka.cs.uiuc.edu/~sondhi1/survey3.pdf> [Retrieved 5 May 2018]
- [42] Spolaôr. N, Monard. M.C, Lee. H.D (2015). *Feature Selection for Multi-Label Learning*. International Joint Conference on Artificial Intelligence. 24: 4401-4402
- [43] Sundqvist. C, Johansson. A.C (2017). *Dialys, peritonealdialys Vårdhandboken*. [Online] Available at: <http://www.vardhandboken.se/Texter/Dialys-peritonealdialys/Oversikt/> [Retrieved 5 May 2018]
- [44] Tomaszewski. W (2012). *Computer-Based medical decision support system based on guidelines, clinical pathways and decision nodes*. Acta of Bioengineering and Biomechanics. 14(1), 107-116
- [45] University of Florida Health (2018). *Home Hemodialysis and Peritoneal Dialysis* [Online] Available at: <https://nephrology.medicine.ufl.edu/patient-care/renal-replacement-therpay/home-dialysis/> [Retrieved 10 May 2018]
- [46] Valderrábano. D, Jofre. R, López-Gómez. M.J (2001). *Quality of life in end-stage renal disease patients*. PlumX Metrics. 38(3), 443-464
- [47] Wedro. B, Stöppler. C.M (2017). *Kidney failure (Symptoms, signs, stages, causes, and treatment)* [Online] Available at: https://www.medicinenet.com/kidney_failure/article.htm [Retrieved 14 February 2018]
- [48] World Health Organization (2011). *International statistical classification of diseases and related health problems (5th ed. 2015)* France. [Online] Available at: http://apps.who.int/classifications/icd10/browse/Content/statichtml/ICD10Volume2_en_2016.pdf?ua=1&ua=1 [Retrieved 13 March 2018]
- [49] World Kidney Day (2015). *Chronic Kidney Disease* [Online] Available at: <http://www.worldkidneyday.org/faqs/chronic-kidney-disease/> [Retrieved 27 March 2018]

ICD-10 Codes

This is a list¹ over the first characters in all ICD-10 codes, standing for the category of the diagnoses.

- A & B: Infectious and Parasitic Diseases
- C: Neoplasms
- D: Neoplasms, Blood, Blood-forming Organs
- E: Endocrine, Nutritional, Metabolic
- F: Mental, and Behavioral Disorders
- G: Nervous System
- H: Eye and Adnexa, Ear and Mastoid Process
- I: Circulatory System
- J: Respiratory System
- K: Digestive System
- L: Skin and Subcutaneous Tissue
- M: Musculoskeletal and Connective Tissue
- N: Genitourinary System
- O: Pregnancy, Childbirth and the Puerperium
- P: Certain Conditions Originating in the Perinatal Period
- Q: Congenital Malformations, Deformations and Chromosomal Abnormalities
- R: Symptoms, Signs and Abnormal Clinical and Lab Findings
- S: Injury, Poisoning, Certain Other Consequences of External Causes
- T: Injury, Poisoning, Certain Other Consequences of External Causes
- U: No coded listed, will be used for emergency code additions
- V,W,X & Y: External Causes of Morbidity
- Z: Factors Influencing Health Status and Contact with Health Services

¹Health Network Solutions (2006). *Anatomy of ICD-10 Codes* [Online]
Available at: <http://www.healthnetworksolutions.net/index.php/understanding-the-icd-10-code-structure> [Retrieved 14 May 2018]

Test Results

This appendix shows all the results from the different tests that have been made during this project. Table B.1-B.4 include the results from the Z-test along with the $\hat{\mu}$ and $\hat{\sigma}$. The Z-test tells whether the test are significantly better, worse or unsure than Baseline. They also include the mean AUC values.

Table B.1: Tests related to the medications group.

Test name	Mean AUC	$\hat{\mu}$	$\hat{\sigma}$	Z-test
M1	0.728629	0.001493	0.000533	Better
M2	0.728924	0.001789	0.000986	Unsure
M3	0.728441	0.001306	0.000818	Unsure
M4	0.7271	-0.000035	0.000685	Unsure
M5	0.727672	0.000536	0.000848	Unsure
M6	0.729192	0.002057	0.001033	Better
M7	0.728616	0.001481	0.000573	Better
M8	0.72862	0.001485	0.000526	Better
M9	0.728248	0.001113	0.000643	Unsure

Table B.2: Test related to the dialysis feature group.

Test name	Mean AUC	$\hat{\mu}$	$\hat{\sigma}$	Z-test
D1	0.727617	0.000481	0.00052	Unsure
D2	0.727791	0.000656	0.000652	Unsure
D3	0.728055	0.000919	0.000707	Unsure
D4	0.728786	0.00165	0.000728	Better
D5	0.728647	0.001511	0.000686	Better
D6	0.728483	0.001348	0.000789	Unsure
D7	0.728018	0.000883	0.0005	Unsure

Table B.3: Tests related to time since visit event.

Test name	Mean AUC	$\hat{\mu}$	$\hat{\sigma}$	Z-test
T1	0.725977	-0.001158	0.000724	Unsure
T2	0.726566	-0.001158	0.000724	Unsure

Table B.4: Test related to the laboratory results and diagnoses group.

Test name	Mean AUC	$\hat{\mu}$	$\hat{\sigma}$	Z-test
L1	0.726769	-0.000367	0.000424	Unsure
L2	0.728116	0.00098	0.000532	Unsure
L3	0.727475	0.00034	0.000472	Unsure
L4	0.720899	-0.006236	0.001084	Worse
L5	0.727086	-0.00005	0.000545	Unsure
L6	0.726201	-0.000934	0.000668	Unsure
L7	0.72681	-0.000325	0.000797	Unsure
L8	0.72882	0.001684	0.000562	Better

Table B.5: Tests related to different combinations.

Test name	Mean AUC	$\hat{\mu}$	$\hat{\sigma}$	Z-test
C1	0.729365	0.002229	0.00096	Better
C2	0.729574	0.002438	0.001099	Better
C3	0.727529	0.000393	0.000925	Unsure

Validation Results

This appendix shows all the results from the validation part. Table C.1-C.4 include the results from the Z-tests along with the mean AUC values. In Table C.4 the Z-test is compared with the old and new feature groups, and in the other tests they are compared together with Baseline.

Table C.1: Tests related to validation part, where balanced data sets were created.

Test explanation	Mean AUC	Z-test
Downsampling		
Test C2	0.726015	Unsure
Baseline	0.728703	Unsure
Upsampling		
Test C2	0.721773	Worse
Baseline	0.719467	Worse

Table C.2: Tests related to the validation part, influence the results.

Test explanation	Mean AUC	Z-test
Test with separated patient groups	1	Better
Test with random feature	0.498636	Worse

Table C.3: Tests related to validation part, where one feature group was removed at each time.

Removed feature group from Baseline	Mean AUC	Z-test
Medication	0.726171	Worse
Dialysis	0.722028	Worse
Time Since healthcare event	0.720215	Worse
Laboratory results and diagnoses	0.718775	Worse

Table C.4: Tests related to single groups were created, both from old feature sets and from the new sets in each feature group. The results from the Z-test is also shown, where the Z-test is based on the difference between the old and new feature groups.

Feature group	Old set mean AUC	New set mean AUC	Z-test
Medication	0.650644	0.665929	Better
Dialysis	0.646770	0.65666	Better
Time since healthcare event	0.655130	0.561418	Worse
Laboratory results and diagnoses	0.702424	0.543892	Worse