Mining for individual patient outcome prediction in hip arthroplasty registry data

Master thesis

Author: Yngve Kristoffersen
Supervisor: Ankica Babic

University of Bergen

The Department of Information Science and Media Studies

1.12.2019

Abstract

The research of this thesis is concerned with developing and evaluating individual patient outcome prediction models based on hip arthroplasty registry data. It was assumed arthroplasty had a rich data collection to be explored using data mining methods. This was conducted in two major phases, firstly exploratory data analysis and then predictive modelling made possible by the finding of the exploration phase. To explore the dataset, clustering was utilized to identify similarities and distinctions between groups of patient records. Resulting from the exploration were the engineering and selection of dependent features to realize the predictive modelling.

The dependent features were used for three separate perspective on modelling a patient outcome grounded in the length of survival of a prosthetic device. These perspectives were two classification tasks with a binary outcome and a multinomial outcome, as well as a prediction of survival as a continuous outcome. The classification tasks attempted to classify patients within categories defined by length of device survival, i.e. above and below eight years, as well as below five, between five and ten, and above ten years. Three separate learning algorithms from *Scikit-learn* were used to examine predictive capabilities in the dataset, and to compare performances. The best performance was observed in the *Multi-layered perceptron* classifier on the binary classification task. The other two algorithms performed comparatively well in binary classification *(Logistic regression* and *Random forest classifier*). None of the models produced reliable results in multinomial classification and in predicting exact survival year. Results suggest that there was not enough explanatory power in the independent variables to perform more complicated predictions.

Acknowledgements

I would especially give thanks to my supervisor Ankica Babic for always being there to help and providing valuable feedback during this project. I would like to thank the Norwegian Arthroplasty Registry for providing the data that made this possible. I would also express gratitude to the occupants of study room 638 for always facilitating interesting and insightful discussions on all things unimportant, and a few peripheral academic conversations. Lastly, a big thank you to my parents for always supporting in times of need.

List of contents

1. Introduction **1**
1.1 Motivation 1
1.2 Research questions 2
1.3 Thesis outline 3
2. Theory **4**
2.1 Orthopaedics 4
2.1.1 Arthroplasty 4
2.2 Data Mining and databases 5
2.2.1 Data Mining in medicine 6
2.2.2 Knowledge Discovery in national registries 7
2.3 Machine learning 7
2.3.1 Pre-processing 8
2.3.2 Unsupervised learning 9
2.3.3 Supervised learning 10
2.4 Related work 13
3. Methodology and methods **15**
3.1 Design Science 15
3.2 Development methodology 18
3.3 Machine learning 18
3.3.1 Data Exploration 18
3.3.2 Pre-processing 18
3.3.3 Unsupervised methods 20
3.3.4 Supervised methods 20
3.3.5 Evaluation metrics 22
4. Technologies and data **25**
4.1 Technologies 25
4.2 Dataset description 26
5. Data exploration **28**
5.1 Approach 28
5.2 Distribution of values 29
5.3 Feature selection 30
5.4 Clustering with K-means 31
5.4.1 Determining number of clusters 32
5.4.2 Survival below five years 33
5.4.3 Survival below ten years 36
5.4.4 Survival below fifteen years 40
5.5 Clustering with Mean Shift 43
5.5.1 Survival below five years 43
5.5.2 Survival below ten years 43
5.5.3 Survival below fifteen years 44

5.6 Summary 44
5.6.1 Features 46
6. Modelling 49
6.1 Approach 49
6.2 Pre-processing 50
6.3 Binary revision classification 52
6.3.1 Patient and device features 52
6.3.2 Patient and primary surgery reason features 53
6.3.3 Patient, device and primary surgery reason features 54
6.4 Multinomial revision classification 56
6.4.1 Patient and device features 56
6.4.2 Patient and primary surgery reason features 58
6.4.3 Patient, device and primary surgery reason features 59
6.5 Predicting exact device survival year 61
6.5.1 Patient and device features 61
6.5.2 Patient and primary surgery reason features 62
6.5.3 Patient, device and primary surgery reason features 63
7. Results 65
7.1 Exploration 65
7.2 Modelling 66
8. Discussion 71
8.1 Methodology and methods 71
8.2 Explanatory power 74
8.3 Challenges and limitations 75
8.4 Answering research questions 75
9. Conclusion 78

Bibliography 81

A List of variables. 85

B List of materials. 87

C List of table header. 88

D K-means below five years result table. 90

E K-means below ten years result table. 92

F K-means below fifteen years result table. 94

G K-means below five years primary surgery reasons tables. 96

H K-means below ten years primary surgery reasons tables. 98

I K-means below fifteen years primary surgery reasons tables. 100

J Mean Shift below five and ten years tables of results. 102

K Mean Shift below fifteen years tables of results. 104

L Table of Mean Shift bandwidth estimation with Silhouette Coefficient. 106

M NSD Approval 107

List of Figures

2.1 Depiction of steps in Knowledge Discover in Databases. 6
2.2 Representation of steps in a standard machine learning pipeline. 9
2.3 Depiction of steps involved in optimizing through cross validation. 12

3.1 Design Science contribution matrix. 16

5.1 Illustration of process taken in exploring by clustering. 29
5.2 Distribution of values in the dataset. 29
5.3 Results from determining number of clusters. 32
5.4 Distribution according to gender below five years, two clusters. 33
5.5 Distribution according to revision below five years, two clusters. 33
5.6 Distribution according to ASA-class below five years, three clusters. 34
5.7 Distribution according to revision below five years, four clusters. 34
5.8 Distribution according to revision below five years, five clusters. 35
5.9 Distribution according to gender below five years, six clusters. 35
5.10 Correlation in primary surgery reasons and revision. 36
5.11 Distribution according to revision below ten years, three clusters. 37
5.12 Distribution according to ASA-class below ten years, three clusters. 37
5.13 Correlation in patient variables below ten years, four clusters, nr.4. 38
5.14 Distribution according revision below ten years, five clusters. 38
5.15 Distribution according to materials below ten years, six clusters. 39
5.16 Distribution according to polyethylene below fifteen years, three clusters. 40
5.17 Correlation in patient variables below fifteen years, four clusters, nr.3. 41
5.18 Correlation in patient variables below fifteen years, five clusters, nr.1. 41
5.19 Correlation in patient variables below fifteen years, six clusters, nr.4. 42
5.20 Correlation in primary surgery reasons below fifteen years. 42
5.21 Steps for engineering binary dependent variables. 46
5.22 Steps for engineering multinomial dependent variable. 46
5.23 Steps for engineering for substitute primary surgery reasons variable. 47
5.24 Distribution of records according to engineered dependent variables. 48

6.1 Illustration of process taken in modelling. 50
6.2 Feature evaluation through Sciki-learn SelectKBest. 51
6.3 Illustration of binary ROC curve, patient and device features. 53
6.4 Illustration of binary ROC curve, patient and primary surgery features. 54
6.5 Illustration of binary ROC curve, all features. 56
6.6 Illustration of multiclass ROC curve, patient and device features. 57
6.7 Illustration of multiclass ROC curve, patient and primary surgery features. 59
6.8 Illustration of multiclass ROC curve, all features. 60
6.9 Regression results line graph, patient and device features. 62
6.10 Regression results line graph, patient and primary reason features. 63

6.11 Regression results line graph, patient and primary reason features. 64

7.1 Illustration of distribution of binary prediction results. 67
7.2 Illustration of confusion matrix for MLP. 67
7.3 Illustration of distribution of multinomial prediction results. 68
7.4 Illustration of distribution of multinomial prediction results. 69

List of Tables

5.1 Silhouette Coefficient scores for clustering with K-means. 44
5.2 Calinski-Harabasz Index scores for clustering with K-means. 44

6.1 Cross validation results with patient and device features (binary). 52
6.2 Cross validation results with patient and surgery reason features (binary). 53
6.3 Cross validation results with all features (binary). 55
6.4 Cross validation results with patient and device features (multinomial). 56
6.5 Cross validation results with patient and surgery reason features (multinomial). 58
6.6 Cross validation results with all features (multinomial). 59
6.7 Cross validation results with patient and device features (regression). 61
6.8 Cross validation results with patient and surgery reason features (regression). 62
6.9 Cross validation result with all features (regression). 64

1. Introduction

The health care sector produces a large quantity of information from a variety of sources which are stored, archived and in the context of quality registries they are curated. This thesis is concerned with Hip Joint Replacement surgery procedures and the inclusion of machine learning practices to search for and build solutions to help predict individual patient outcomes. In Scandinavia and other parts of Europe the maintenance of registries in hip and knee arthroplasty has been on-going for several decades (Delaunay, 2014). When a surgery is performed, either primary surgery or revision surgery, information deemed important are recorded and later entered into national spanning databases. This practice of data accumulation establishes potential for exploration and acquisition of new knowledge and solutions. Hip joint replacement surgery is most commonly associated with the elderly in society, although some change has occurred as it is becoming a more frequent phenomenon even in younger patients. The elderly part of the population in contemporary western society is additionally increasing in size, a corresponding increase in the necessity of primary hip joint replacement surgery is observed (Furnes, 2019). Registry data have been utilized for detecting surgery factors relating to survival rate of patients and for examining risks and reasons for requiring revision surgery (Varnum, 2019). Furthermore, it can provide specific information on surgery and products used in joint replacement, as well as, providing benchmarks on performance of specific prosthetic devices (Varnum, 2019).

1.1 Motivation

The *purpose* in this research is to explore and attempt to better understand the explanatory potential in registry data collected on surgery cases for hip joint replacement. Furthermore, to assess implementation of machine learning algorithm on predicting and classifying potential outcomes for individual patients.

The *reason* for performing such research is to see whether existing features available in the registry could be used to better understand the outcome of implanting a specific prosthetic device. Focusing on features know to clinicians and researchers before carrying out the primary surgery. The research is largely consistent of two parts, firstly the data is explored to assess what dependent and independent features are present, and if outcomes exist or can be engineered by merging existent features. The second part involves modelling learning algorithms on the features selected from the exploration to assess the potential of predicting individual patient outcomes. As the number of primary surgeries increase within the population, the necessity for stable and relevant solutions become more significant. In machine learning a large part of the potential to predict arrive from the explanatory power of independent features and the correct choice of learning algorithms (Buitinck et al, 2013). Therefore, this research explores the possibility to use registry data to predict individual patient outcomes by applying data mining using open source software for machine learning.

1.2 Research questions

Q1: Which variables in the dataset are suitable as dependent outcome features in this excerpt from a quality registry on hip arthroplasty?

Q2: Which variables in the dataset have potential as independent features for explaining an outcome after hip replacement surgery?

Q3: Can the dataset and a selection of learning algorithms give reliable results in predicting an individual patient outcome?

The research questions address three separate parts of the overall exploration and evaluation of potential in variables and use of learning algorithms on the dataset. The first question addresses the potential for locating features denoting outcomes which can occur after having implanted a prosthetic device, and/or can be used to engineer dependent variables. The second question address the issue of finding the features for explaining whether a patient will have a certain type of outcome. Lastly, the third question addresses an attempt to determine which outcome will happen by applying algorithms to learn from the relationship between target (outcome) and explanatory features.

Overall, the research questions address potential for different individual patient outcomes described by variables available in a registry-based dataset. The feasibility of methods designed by training on separate explanatory features are evaluated by metrics. Furthermore, together the questions investigate the potential for building adequate solution by leveraging registry data in hip arthroplasty.

1.3 Thesis outline

The outline of the thesis is listed below, excluding this chapter.

Chapter 2 – Theory: This chapter presents the theoretical foundation, and discuss medical informatics, orthopaedics, knowledge discovery and machine learning. As well as related research performed on similar data for a purpose akin to this thesis.

Chapter 3 – Methodology and methods: This chapter discusses the methodology guiding the research, and the methods used for performing exploration, modelling, and evaluation of the results.

Chapter 4 – Technologies and data: This chapter introduces the dataset used in this thesis, a brief explanation of contents, and technologies necessary to utilize the chosen methods.

Chapter 5 – Exploratory data analysis: This chapter is a walkthrough of the exploratory data analysis by clustering, and selection of outcome features as potential perspectives on predicting an individual patient outcome.

Chapter 6 – Modelling: This chapter details the modelling tasks performed for the selected outcome features from the exploratory data analysis.

Chapter 7 – Evaluation: This chapter evaluates the performance of classification and regression algorithms implemented in Chapter 6.

Chapter 8 – Discussion: This chapter deliberates on the results from exploration and modelling, and the potentials and issues encountered throughout this research.

Chapter 9 – Conclusion and future work: This chapter sums up the findings in this thesis, as well as advice for how to conduct further research.

2. Theory

This chapter details the literature related to the purpose of the thesis, and discuss medical informatics, orthopaedics, knowledge discovery and machine learning. As well as related research performed on similar data for a purpose akin to this thesis.

Medical informatics pursue to fill a gap created by the unification of medical science and its many dimensions, and the development of information systems. The process of delivering medical and healthcare service by utilizing computer-assisted methods available through development of new methods and equipment. Services developed for the health care sector may attempt at assisting medical personnel in their daily routine, simplifying communication, and aid in making decisions for both patients and practitioners. Medical informatics relies thereby on theory from information sciences alongside medical sciences for assisting in managerial tasks, diagnosis, and treatment by employing resources, methods and devices to gather, store, retrieve, and utilize information to maintain and improve standards of practice (Closa et al, 2009, p. 155).

2.1 Orthopaedics

Orthopaedics is concerned with the human muscle and skeletal system. The bone is connective tissue made of both organic and inorganic matter, and the consistency of our bones is altered as we grow older (Iyer, 2013, p. 2). The change in bone caused by aging or by any adverse event, such a fracturing, can lead to the immobility in parts of the body (Iyer, 2013, p. 405). The skeletal system is intricately connected to the muscles and blood vessels in our physiology and how we as persons move our body parts. Disruptions in how the skeletal system functions can therefore cause larger problems of pain and trauma, some which requires surgical treatment (Iyer, 2013, p. 17).

2.1.1 Arthroplasty

Joint replacement is the process of removing our natural joint when the articular surface has been deteriorating by arthritis, common reference to multiple illnesses affecting the joints, or by fracturing the joint. Arthroplasty is another term used for referring to joint replacement or realignment, the main goal when performing arthroplasty is to relieve of suffering and restore functionality (Iyer, 2013, p. 317). There are several different types of arthroplasty, total or partial joint replacement, resurfacing arthroplasty, excisional arthroplasty, and interposition arthroplasty. Most common areas where arthroplasty is performed is in the hip, knee and shoulder, and may also be in less common areas such as ankle, elbow and wrist (Iyer, 2013, p. 317-319).

Common reasons for joint replacement

Common causes for when joint replacement is an appropriate action, to mention a few, are: *coxarthrosis*, a breakdown in the surface material on joints and the underlying bone, *rheumatoid arthritis*, an autoimmune disorder causing pain and damage on the joints, *post-traumatic arthritis*, a form of arthritis occurring post injury due to the effect of said injury, *avascular necrosis*, a disturbed joint cause blood flow interruption and tissue degeneration. Iyer (2013, p. 228-330) also discuss a few other events that may result in failure and cause of revision surgery, *aseptic loosening of stem and/or cup*, *infections*, *wear on parts of the prosthesis device*, and *infections*. A national report from the Norwegian registry details a majority *coxarthrosis* as the cause for primary surgery, and an increase in number of required surgeries in later years, although the revision rate is stated as the lowest registered rate in its history at 12.7% in 2018 (Furnes, 2019).

Total hip arthroplasty

Total hip arthroplasty (THA) is the surgical procedure of implanting a prosthetic joint after removing an arthritic or fractured joint. The purpose is to improve and regain function, as well as relieving a patient of pain. A prosthetic can be of materials such as metal, ceramics, and polyethylene. A modern prosthetic device for THA *consists of a femoral stem*, *femoral head*, *acetabular shell* and *acetabular liner* (Iyer, 2013, p. 326-328). There are different practices with the use of prosthetics that have impact on the patient, for example, a larger head size on the prosthesis can cause less chance of dislocations and more wear while a smaller size can cause higher probability of dislocations (Iyer, 2013, p. 328). Another aspect is the resurfacing of the artificial head; with the use of metal prosthetics an increased amount of metal ions has been reported, but with unknown risks. The survivorship of revision is high, but there are occurrences of premature failures in some series of prosthetics devices (Iyer, 2013, p. 329). Iyer (2013, p. 334) also notes *Revision Hip Joint Arthroplasty* as a technically difficult procedure.

2.2 Data Mining and databases

Data mining is the task of exploring data to uncover new information, it is described as having acquired its foundational methodologies from three fields, statistics, computational methods, and data visualization (Gorunescu, 2012, p. 2-3). The methods used to *"mine"* data is an approach suitable to most domains, from health care to finance and energy related data. Databases have by necessity and possibility been established in all these domains, ranging from unstructured data in vast cloudy databases to more refined quality registries (Sarkar, p. 48). The focus here is specifically within the domain of health care and medicine. Presently we record data in large quantities, a phenomenon which have during the last decades increased at an excessive rate due to development of new technologies adopted by the medical sector, as well as the diversity of other private and public domains (Hilbert & Lopez, 2011), (Gorunescu, 2012, p. 5-6).

The data is firstly prepared, cleaned and transformed in some manner, important features are looked for by evaluation metrics, and by unsupervised methods from machine learning, as well as by visual exploration (Figure 2.1) (Gorunescu, 2012, p. 6).

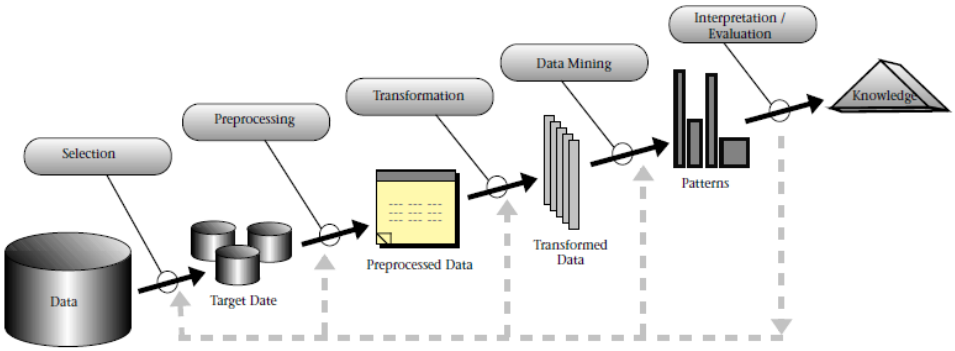

Figure 2.1: Depiction of the steps from the KDD process (Fayyad & Uthurusamy, 1996).

2.2.1 Data Mining in medicine

Databases in health care have by now a presence in areas such as effectiveness of a treatment and understanding reasons for occurring symptoms, assisting with decision-making for patients and clinicians, and detecting fraudulent behaviour (Koh and Tan, 2005). Data from a range of domains have been investigated, clinical data from patient with ADHD have been tested by applying machine learning methods. Other promising attempts are based on raw biomedical data to predict cancers outcome, as well as exploring significant factors for survival of patients diagnosed with end-stage kidney disease (Yoo et al. 2012). Raw data rarely give beneficial information directly without attempting to uncover any, automated data analysis by statistical and computational methods adopted by machine learning provides the methods to perform what has been dubbed Knowledge-discovery in Databases (KDD) by *mining* (Fayyad and Uthurusamy, 1996).

The development of the capacity to utilize large quantities of data alongside data mining techniques have increased community awareness of data analysis for finding new information. Mackinnon and Glick (1999) mention Chatfield's' definition of data mining:

> *"the extraction of previously unknown information from databases that may be large, noisy, and have missing data"* (Chatfield, 1997, cited in Mackinnon and Glick, 1999)

Chatfield's definition includes a few important points on data analysis, how it can be noisy and have missing data, and needs pre-processing. Gorunescu (2012, p. 57-58) elaborate on a

process in data mining referred to as *exploratory data analysis*. A first step of search and retrieval relying on the ability to describe and locate correlations, patterns or phenomena, and possibly extract important features (Gorunescu, 2012, p. 57). Large and noisy databases may at first appear lacking in knowledge, to make them become fruitful they often has to be subjected to analysis and interpretation (Obermeyer and Emanuel, 2016).

2.2.2 Knowledge Discovery in national registries

Registries in Orthopaedics have been established in several countries now, a registry is a comprehensive collection of data on one or several health care related aspects pertaining to a defined population (Delaunay, 2014). Scandinavian countries have for the latter two decades or so been maintaining registries with data on conditions relating to the musculoskeletal system, such as the Norwegian Arthroplasty Registry of THA, Danish Hip Arthroplasty Registry, as well as the nationwide Canadian Joint Replacement Registry and the New Zealand Joint Registry for THA and TKA in other countries (Delaunay, 2014). Establishing registries have supplemented medical practitioners with the ability to compare the issues encountered within their work with that of others in the same speciality, and on a nation-wide scale.

Machine learning has become an important aspect of KDD, relying upon algorithms for assisting in knowledge engineering, problem solving by finding best optimizations of an algorithm on a dataset, and for performing a prediction or classification (Mackinnon and Glick., 1999). There are also the possibilities of combining quality registry data on Hip Arthroplasty with other health data registries to acquire the data to build models valid at an individual level. The main purpose is to get access to the necessary fine-grained data to allow for constructing algorithms suitable for implementing shared decision-making systems for patients who are contemplating Total Hip Arthroplasty (Cnudde et al, 2016). The value in quality registries are discussed as being not fully utilized, but said to hold potential for effective services to be constructed and serve as a high value advantage which could be beneficial for patients and clinicians (Nelson et al, 2016).

2.3 Machine learning

Machine learning deals with how computers can learn to recognize patterns by processing data and examining relationships between the data. Arthur Samuel gave a less technical definition on what machine learning is:

"*Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959, in Geron, 2017, p. 4)

Machine learning has become more popular with the development of faster computers and access to a wider range of data in volume and variety (Yoo et al, 2012). The aspect of not having to program the rules the algorithm bases its reasoning on makes machine learning require less explicit programming to run, optimize and test, and more easily adoptable on a broader scale. As well as making it a more ideal solution toward problems with a large amount of data or handling varieties in data which can be an excessive amount of work for persons (Geron, 2017, p. 6-7). A more technical explanation on what is machine learning is given by Tom Mitchell:

> *"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."* (Tom Mitchell, 1997, in Geron, 2017, p.4)

There are three reoccurring variables throughout the above definition, the performance P at fulfilling a given task T which improves through iterations from rendered experience E. Machine learning in a brief statement is learning algorithms which execute the functionality described above (Sarkar, p. 10).

2.3.1. Pre-processing

Pre-processing data is necessary to transform it to a form suitable for the learning algorithm to ingest, since datasets often have features understood through separate scales, noise and missing values (Butinick., 2013).

Dimensionality reduction

Voluminous data with multiple differences in nature and variety to consider can induce an increase in problems complexity, causing the precision of an outcome to decrease (Sarkar et al, 2018, p. 39-40). In simple terms, the smaller the dataset the less complicated it is to analyse it, and the other way around, but this is regarding only an absence of complication not what it gives back in rewards. Dimensionality reduction can be performed through two approaches to the same problem of narrowing the number of features to consider, *feature selection* and *feature extraction* (Sarkar et al., 2018, p. 39-40). The importance of reducing dimensions can be understood through what has been dubbed '*The curse of dimensionality'*. The *curse* refers to the phenomenon that arises in situations where the purpose is to analyse a quantity in variety, as the number of dimensions increases, the corresponding feature space becomes larger (Sarkar et al., 2018, p. 40). This can cause pieces of actual importance to become scattered more thinly and harder to observe.

Selection and engineering

Through *selection*, the feature to proceed with in further analysis and modelling is selected from existent features in the dataset. Meanwhile, *feature engineering* is done by combining existing features into new ones by imposing conditions on the data as explanations of an observed phenomenon. Rows in a dataset are then labelled by these conditions (Sarkar et al, 2018, p. 53).

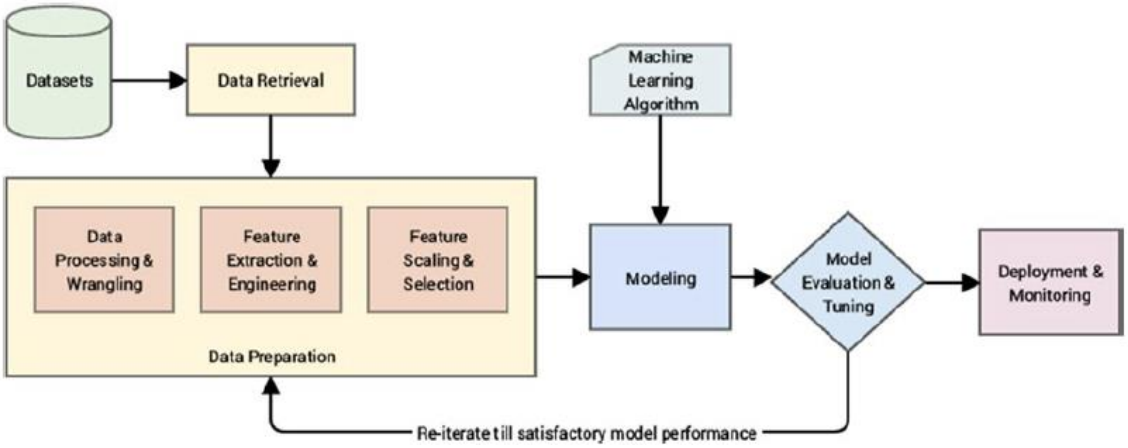

Figure 2.2: Depiction of standard machine learning pipeline (Sarkar, 2018, p. 53).

Standardization

Standardization is the process of reducing the value in a feature vector with attributes from different natures, and therefore represented by separate measurements, to be given on a similar scale. To illustrate, a feature vector on a surgery case may hold a specification on device size in centimetres all below a hundred in diameter, while another representing biometric data represented in the thousands. This difference in the size and nature of the scale can cause confusion (Sarkar et al, 2018, p. 180). Standardization resolves this by calculating the mean and standard deviation for an attribute, then further subtracting the mean and split it on deviation for each value of that attribute (Geron, 2017, p. 66-67).

2.3.2 Unsupervised Learning

Unsupervised learning is concerned with data that often have no known ground truth available to tell the conditions of the data and about phenomena it holds. Regarding the definition of the process of learning an algorithm, unsupervised methods gain the experience E in working without the advantage of pre-labelled data to interpret the phenomena. Rendering it as a powerful means to analyse data to find insight into what trends or patterns it may yield for understanding more about its nature and what it describes (Sarkar, 2018, p.39).

Computational methods in unsupervised learning do not take any guidance in deriving values from the data and assign them by the use of metrics, therefore it is not given previous experience in performing the given task. Instead it does the heavy lifting by itself as it attempts to find inherent latent structures and relationships between data points. Broadly defined through several different sub-categories, such as clustering, dimensionality reduction, anomaly detection and association rule mining (Sarkar et al, 2018, p. 40).
Unsupervised learning is often used prior to predicting or classifying and can be used for exploring data points which should have similar labels by for example indicating a relationship in outcome similarity or as risk groups. A common method for grouping and assigning labels is different types of clustering algorithms (Sarkar et al, 2018, p. 39).

Clustering

Clustering can be performed by using several different methods, the idea is to group data by deriving an assumed similarity in and relationship between data points, resulting in defined clusters/groups. Clustering handles data without any prior training or already known contextual knowledge about the data points, then produces a label for each data point which is retrievable after processing (Sarkar et al, 2018, p. 260). Resulting formations made up by each data point's assigned membership to a range of groups is subject for interpretation and evaluated by internal or external metrics on distance and density (Sarkar et al, 2018, p. 279-280).

Partition-based clustering approaches the problem by establishing a notion of similarity that is defined through applying mathematical function on data points (Sarkar et al, 2018, p. 260). The measurement of similarity is further used to separate data into groups by starting at a frivolously chosen attribute and comparing and reassigning until each reassigning does no significant change to the distribution. K-means is one example of a partitioning based cluster and is commonly used for data with spherical formations (Sarkar et al., 2018, p. 260).

Density-based clustering approaches the problem in a different way, giving up the notion of *distance*, and rather defines a notion of *density* as a way of handling arbitrary shaped clusters. The clusters are formed by finding areas with greater quantities of data and works well as it is an unlikely event that all detectable clusters are spherical in nature (Sarkar et al, 2018, p. 260).

2.3.3 Supervised Learning

Methods from supervised learning focus on mapping the input data passed into the algorithm to a corresponding output by examining a record of inputs and subsequent outputs that have been set aside, often referred to as a training set. This is done to train an algorithm to attempt to understand how the inputs and related outputs are associated and create a trained algorithm. It is then used to predict an outcome by running the model on previously unobserved data.

The training is an attempt to model the relationships present in the training data, and take this knowledge gained and reuse it to predict. In contrast to unsupervised methods for machine learning, supervised methods must take a record of outputs that corresponds to an input, and the relationship between these acts as guidance on how to interpret new incoming values after an algorithm has learned from experience (Sarkar et al, 2018, p. 179).

Considering the definition of machine learning, the experience E is gained in a controlled environment in supervised learning, as a means to prepare and optimize in performing task T, an option the unsupervised solutions do not have.

Generalization

Generalization in a learning algorithms performance is how well it does on newly observed data not present in the training set. Poor generalization can be an issue as the algorithms cannot be applied to a larger set of samples from its specific area. A goal in learning is for the model to be applicable to a broader range of sample without either being to specifically or to loosely tuned to the relationships in the training data (Sarkar, 2018, p. 287).

A fitting issue arise when pursuing good standard of generalization across the spectrum of different data and can be understood through the trade-off between *bias* and *variance*. *Bias* refers to the model's competence in making the right decision, and measures the error rate, or deviation, between what the truth was and what it was assumed to be. A high bias is related to larger presence of noise (Sarkar, 2018, p. 284-285). This causes the model to miss out on learning how to make the correct assumptions about relationships and eventually it makes erroneous decisions. *Variance* refers to the range of difference in performance across a changing set of data samples; low variance implies stability in performance, while high variance implies a larger difference in error rate in predicted outcomes (Sarkar, 2018, p. 284-285). The issue causes a problematic situation in building a model that adopts to change and still renders reliable results without being misled by newly discovered features, noise, or randomness. *Under-* and *Overfitting* are two scenarios to consider to better understand the occurrence of high or low bias and variance in modelling a problem space.

Underfitting is when the model is incapable of learning anything from the underlying structure, patterns and correlations in the dataset, and is characterized by low variety in performance and high bias as the model made no clear assumptions about relationships in the data (Sarkar, 2018, p. 287). A model with these attributes will then have a stable range of outcome predictions as variety is low, although there is stability in variety, the algorithm makes frequent mistakes.

Overfitting is the somewhat different in trade-off between *variance* and *bias*. With high bias, as the wrong assumptions is made on structures and patterns in the data, and high variance, as the noise and randomness in the data is assumed as informative and influence the

predictions. This is caused by too strictly fit assumptions toward the sample of data used for training and can generate lots of errors in modelling unobserved data (Sarkar, 2018, p. 287).

Optimization

Optimization involves a set of approaches toward preventing overfitting and is often performed by *cross validating* with a selection of hyperparameter settings (Figure 2.3). The method combines the settings in all possible combinations and return the one with the best performance (Varoquaux et al, 2017).

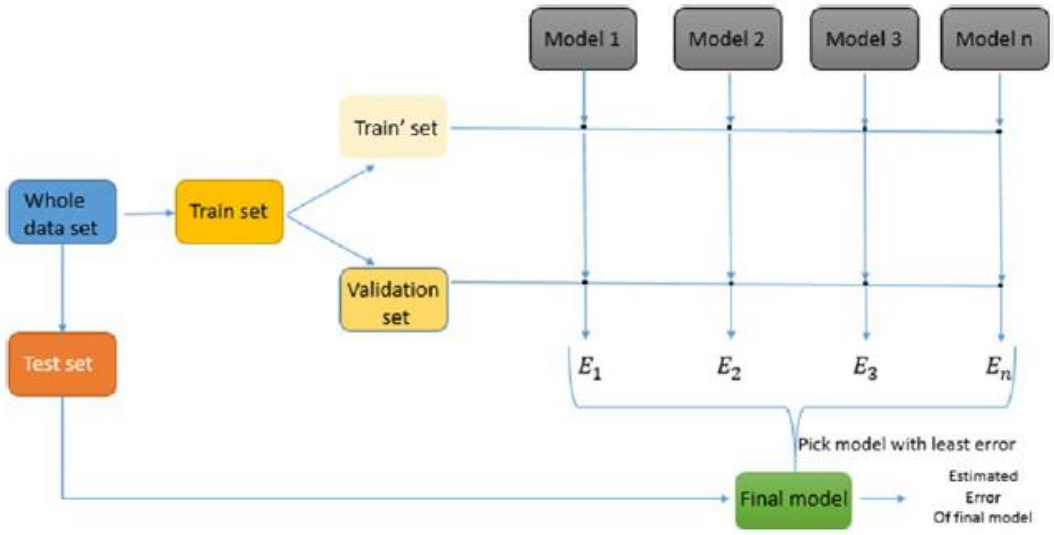

Figure 2.3: Depiction of cross validation to select best model by measuring least error (Sarkar, 2018, p. 289).

Regularization refers to finding a suitable method for reducing the chance of overfitting in a algorithm's learning process by reducing the complexity. *Ridge regression* is a common form of regularization done by putting a constraint on the coefficient, adding bias but reducing variance in the resulting outcomes (Geron, 2017, p. 127-128).

Stochastic gradient descent can be applied through hyperparameter settings in *Multi-layered perceptron* classifier from Scikit-learn (Varoquaux et al, 2017). The literal meaning of *gradient descent* explains the nature of the method in very simple terms, as it picks a starting position and *moves downward* the slope of a cost function in iterations for each feature until it gets as far as the lowest point. *Stochastic gradient descent* is one of the more common methods for optimizing learning of a model by adjusting the rate it moves downward this slope, or size of previous update, and tracking it to improve performance through the learning process (Geron, 2017, p. 117).

Regression

Regression refers to tasks predicting a continuous numeric value or a class in case of logistic regression. This is performed by estimation through mapping the relationship between input features and corresponding output from a dataset with prepared target feature and explanatory features (Sarkar et al., 2018, p. 37). A common example where regression would be the choice is the prediction of housing prices, where input features such as size, age, amount of bedrooms/baths etc. is passed in to guess what the price will progress to from its' current value, i.e. the price of a house by next the next quarter of the year in contrast to its current value and the values stored from the previous quarters.

Classification

Classification refers to all tasks that attempt to predict an output that is categorical in nature as it pertains to one out of a set of distinct classes, usually a fixed number of available classes. Classification can further be divided into binomial and multinomial classification, where the difference between the two is the amount of available output classes (Sarkar et al., 2018, p. 36). If the prediction can be a good or a bad result it is binomial, while the prediction of what genre a book or movie belongs to is a multinomial classification due to the existence of several classes of possible outcome.

2.4 Related work

This section is a review of literature relating to the area of interest in this thesis, the combination of patient outcomes in orthopaedics and machine learning methods. It is comprised as a discussion of similar research where data has been assessed and utilized for predictive modelling.

Kruse et al. (2017) discuss applying machine learning methods to hip fractures patients in X-ray data from Denmark in combination with additional data from a period of five years after surgery. Their purpose was to find patients who might sustain fractures from osteoporosis-related reasons and classify those who would or would not be in risk, and reported positive results measuring an *area under the curve* score above 0.80. The data originated from national Danish patient data and collection of images from two university hospitals on patients with records on hip and/or femur region fractures. They suggest results from the study can be improved by supplementary data from a larger region and could be beneficial for identifying patients with a certain risk (Kruse et al., 2017). The chance of improvement in additional data is something also emphasized by others who suggest implementing solutions for predicting individual patient outcomes (Ellison, 2017).

Fontana et al. (2019) reported on a similar venture, relying on patient reported outcome measures from Total joint replacement surgery and a set of supervised machine learning methods. They used *Logistic regression*, *Random forest* decision tree classification and *Support*

vector machine (SVM) to predict whether a patient would show changes that are of *minimally clinically important difference,* approaching the issue as a binary classification.

Their results are discussed as being in-between poor and good, with a variation in area under the curve scores ranging from 0.60s to the 0.80s (Fontana et al., 2019). The performance was improved from supplementing additional patient reported outcome measures, however, it is discussed as not improving by adding extra data from hospitalization (Fontana et al., 2019).

The issue of providing informative solutions as clinical aid by predicting future outcomes are present in literature, a summary of several results discussed in an article reports uses of *Logistic regression classifier* and *Random forest classifier* with decision trees, and approaches the problem based on rotating independent features for testing performances (Cabitza et al, 2018). Further, several attempts at clustering data are discussed, for reasons such as segregating records into groups to identify patients with different levels of risk of suffering an adverse event after surgery and locating fracture risks in patients (Cabitza et al, 2018). Approaching the is issue of finding groups with similar characteristics to compare and establish an understanding of the data can be done by model-based clustering (Fraley and Raftery, 2002). Investigating data in hip replacement using larger registries has for instance been done to locate reasons for differences in efficiency of revision surgery among patient records (Salassa et al, 2014). The investigation is useful for designing better predictive models, as their performance is a result of design by the training data.

Assessment tools for outcome predictions have also been tested for a variety of causes relating to arthroplasty surgery; such as predicting postoperative rehabilitation needs, in-hospital care needs after surgery, and occurrence of postoperative complications (Konopka, 2015). A variety of variables are discussed throughout the article, most focused on general patient information, i.e. among others age, gender, BMI, and health status by the American Society of Anaesthesiologists classification system, in combination with a selection of other contextual variables (Konopka, 2015). The possibility of building larger data warehouse structures has led to the collection of quantities of data in medicine. However, even as registries of records exist, having the right data is an issue, introducing a trade-off in quantity and quality, with one not being enough. Volume is necessary to make sure enough entities are present to represent the true associations in the data, and not a reflection of only a small part of the population (Roski et al, 2014), as using inadequate knowledge may cause poor decision-making. Quality is necessary as even though a dataset might be large, it may also be noisy and have randomness misrepresenting the true associations, causing the model not to learn what it needs to make good decisions (Roski et al, 2014).

The accumulation of data on patients in arthroplasty and the increased availability of machine learning methods have caused a surge in research and development of shared decisions making solutions (Bozic, 2013), (Nemes, 2018). For instance, such solutions could aid clinicians and patients in making better decisions together by increasing awareness of outcome possibilities. However, to realize this the issue of the right assembly of data and validated methods to implement is a fundamental necessity.

3. Methodology and methods

This chapter discusses the underlying methodology guiding the research, and methods used to perform the necessary actions and evaluations through the process of analysing records and predictive modelling for patient outcomes. The evaluation in this thesis is done by applying performance evaluation metrics to assess the feasibility of the results from unsupervised and supervised machine learning methods.

3.1 Design Science

The development of new systems in information technology is concerned with the idea of improving efficiency and contribution to the organizational structure or a determined environment. New systems can often be complex, intact with advancements in technological capabilities, there emerges a necessity for studying such systems on multiple planes. It is therefore argued that to approach the problem of researching information systems requires the inclusion of two paradigms of science, *Behavioural* and *Design Science* (Hevner et al, 2004). The design science paradigm is a problem-solving centric methodology, rooted in the engineering field, and attempts to construct artefacts in a research context by applying related knowledge in the process to ensure relevance and rigor in research (March and Smith., 1995). Behavioural science is more concerned with providing explanation to questions about why and how a phenomenon is the way it is, seeking out some truth by understanding and conceptualizing how things work (Hevner et al, 2014). Regarding the construction of an artefact within information technology, the behavioural science paradigm provides a sense of direction by promoting the development and justification of more precise theories for explaining how the artefact should be shaped (March and Smith, 1995). For designing systems, theories can often relate to the performance of a system in an organizational structure, in efficiency and/or achieving usefulness in executing desired functionality (Hevner et al, 2014).

Design science as a methodology is different from the concept of development methodology, while the latter focuses on best practices for maximizing efficiency in structure and management of a development process. The first is more concerned with the novelty solutions within a research perspective. The goal is to prescribe solutions toward problems with the intent of either improve upon a previous solution or address unsolved issues through an innovative approach. Likewise, Design Science research helps encourage contribution towards a knowledge base (Gregor and Hevner, 2013).

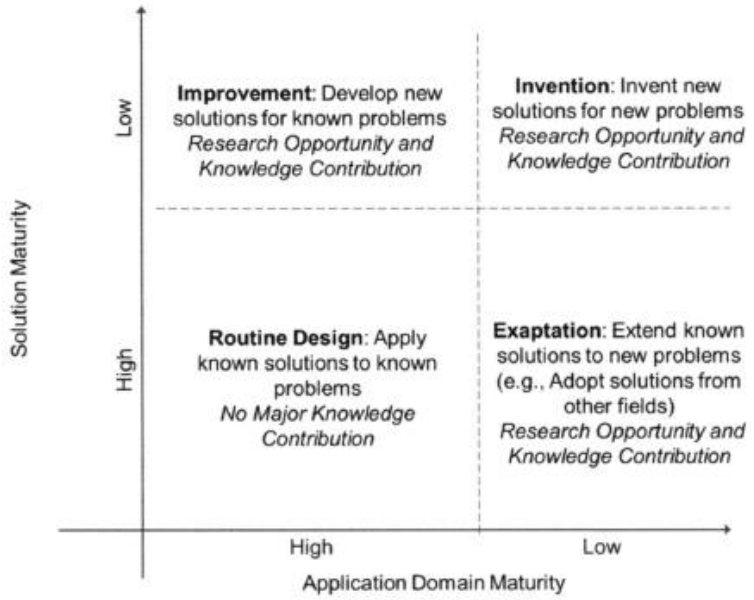

Figure 3.1: The Design Science Research contribution matrix (Gregor and Hevner, 2013)

The design science process: In behavioural sciences the focus is on constructing theories on how things work and evaluating the results. Correspondingly, the design science process is involved with activities generalizable to constructing artefacts of some kind and evaluating the result (March and Smith, 1995). Artefacts are discussed as belonging to a list of categories detailing types of products:

- Constructs are artefacts which assist the communication of knowledge within a domain by aiding the composition of a shared vocabulary (March and Smith., 1995). An example of a construct could be attributes and entities as representations of a certain concept.

- Models are representations of the real-world by relying on constructs to build composite structures based on their relationships (March and Smith., 1995). An example of a type of model could be an Entity-Relationship diagram illustrating which attributes are involved in the assembly of a representation of a phenomenon.

- Methods are the means for conducting an activity aimed at performing a certain goal (Hevner et al, 2014). An example of methods would be the ways to perform certain activities by providing descriptive textual approach or computational algorithms as the goal.

- Instantiations are the step of bringing all other types of artefacts, from attributes at the bottom level to validated methods at the top, together to realize a system in its true environment (March and Smith., 1995). Implementations of instantiations are done to illustrate how adapt an artefact performs in a real-world context and to establish proof its usefulness.

Evaluation: The evaluation is often concerned with testing adequacy in performance and benefits rendered by an artefact to address a problem and provide solutions (Hevner et al., 2004). Sufficiently evaluating research can lead to providing better information on the results and therefore highlight potential for improvement in further research (Hevner et al., 2004). Evaluating the performance of an artefact is a relative exercise as it is highly connected to the nature of the artefact, and/or the intended functionality. Which metrics are used to perform an evaluation often depend on the environment of a given artefact and its intended usage and can range from assessment of completeness and functionality to usability and reliability. As well as how technological implementations are used within desired environment (Hevner et al., 2004).

The study of information systems involves two different types of research, descriptive and prescriptive. The first is concerned with providing adequate knowledge about the intended problem space and related phenomena to avoid poorly constructed solutions. The latter is concerned with identification and investigation of artefacts similar to the intended research, assisting in restricting domain knowledge to a baseline and preventing over-reaching (Hevner et al, 2004).

The Artefact

The artefact in this thesis is in form of engineered and selected dependent features for suggesting possible predictive solutions for hip arthroplasty patients. Furthermore, predictive models are produced to test different outcome prediction possibilities as a suggestion for further use of machine learning in utilizing registry data to benefit clinicians and patients alike. The evaluation of the artefact was performed by evaluation metrics, such as the AUC score to document achieved performance (Hajian-Tilaki, 2013).

The artefact consists of models and methods, as defined above, the attributes of registry records are explored to synthesize the constructs into applicable models for interpreting the data, i.e. defining new features based on the attributes in the dataset. Further, these models are used to build methods to test performance of a goal-directed activity, i.e. to predict/classify an individual patient outcome. Generally speaking, the artefact is in two parts. First part concerns locating and/or establishing the outcome features for describing a distinct difference in outcomes between patients. The second part is implementing learning algorithm to perform the goal-directed activity based on the outcome features established in the first part.

3.2 Development methodology

The development of the artefact in this research project was performed through exploration of the problem space established by the research questions. This included data analysis with unknown results and therefore found a methodology with the ability to adopt quickly changes in development. Crystal Clear is an agile methodology with a focus on creating room for navigation during the process of development (Cockburn, 2014). The methodology suggests frequent deliveries and communication, and reflective improvement through clarification and evaluation (Cockburn, 2004). The methodology's impression of adaptiveness to a changing environment made it appear suitable for this thesis.

The development was done in two iterations. The first corresponds to the phase of data exploration and feature engineering, the second iteration was concerned with predicting individual patient outcomes and evaluating results. The machine learning models used to perform the necessary activities to realize this are presented next.

3.3 Machine learning methods

The methods from the field of machine learning used to pre-process, explore and model are described in more detail. Included are the methods for unsupervised clustering, regression and classification, as well as, an overview on pre-processing the data.

3.3.1 Exploratory Data Analysis

Exploratory data analysis focuses on interpreting and establishing understanding about a dataset, recognizing systematically underlaying trends and patterns. Often this exploration can be aided by use of different techniques for identifying the systematic relationships between attributes/variables present in the data (Gorunescu, 2011, p. 57). Exploring was chosen as method for revealing the underlying structures, detect differences, and maximizing knowledge about what the records detail about hip replacement surgery. Furthermore, to identify important variables within the data, and suggest potential outcome features.

3.3.2 Pre-processing

Pre-processing the dataset used in this thesis was done to make it appropriate in shape and form for the learning process, the step of transforming the data before use is done prior to exploration and prior to modelling in this thesis. Description of pre-processing methods are following.

Feature selection

Feature selection is the act of deciding which of the attributes in a dataset to include, it is done twice throughout this thesis, prior to exploration and prior to modelling. Variables are also assessed while exploring the dataset to decide other possible dependent and independent features. The process of selection was guided by the established purpose of the research, focusing on locating descriptions of an outcome and the means to predict said outcome after primary surgery.

Feature engineering

Feature engineering refers to the act of establishing new features from existent ones within the data, to better describe the structures and phenomena within the dataset. The creation of new features is often a demanding task as it can require necessary knowledge to direct the process. However, it is a highly important aspect as it can be used for establishing dependent variables or reducing the number of dimensions by combining independent features (Sarkar et al, 2018, p. 181-182). Engineering new features in this thesis was performed as a solution to establish more relatable outcome categories after exploratory data analysis. Furthermore, to check for improvement in performance with different dependent features.

Binary encoding features

Numerous features from the dataset are in a categorical multinomial range without any order or context to help the algorithm with interpretation. For portraying these values in a way to minimize the machine learning models chance of learning erroneous associations, *one-hot encoding* was used to transform the multinomial independent features to a binary matrix representation (Buitinck, 2013). For instance, *prosthetic device materials* are represented as a defined numerical scale, but the materials represented by the values three and four are not any closer associated than three and ten.

Scaling features

Feature scaling is performed on continuous features selected from the dataset, by reducing them to fit on the same scale. To illustrate, features may have different nature, a person’s age and their weight are both numerical, but describe different phenomena and may have larger difference in the size of the scale. This difference in nature can, if not tended to, cause a decrease in performance (Geron, 2017, p. 65). Standardization was utilized as the method for scaling values in this thesis.

3.3.3 Unsupervised methods

Clustering models were used for exploring the data in the attempt to find possible outcomes in groups in the dataset. To perform this action *K-Means* is used to group cases on similarity between cases as a method for examining the population. *Mean Shift,* based on a different internal strategy for clustering is used as a secondary method for comparison. Both were provided by the Scikit-learn environment (Varoquaux et al, 2017).

K-Means

K-means belongs to the partitioning family of machine learning models but differs from other clustering models in the ability to manually adjust the K number of clusters the data should be grouped into as a result (Sarkar et al., 2018, p. 386). This gives a unique opportunity to define a space by setting thresholds of K clusters. The metric used to measure distance is *Euclidian distance*, measuring the relationship between two data points in a set. It is used to decide where to position data points towards the cluster with the centroid it is nearest, and performs this for each data point. The centre is recalculated by averaging the dimensions among data in each cluster and moving data points to the closest group. After all data has been assigned then the process repeats itself iteratively, the iterations are repeated until cluster formations are stable (Sarkar, et al. p. 267).

Mean Shift

Mean Shift is a density-based cluster, although it is similar in use of centroids for defining a search space, it measures where the denser areas are by *shifting* across the data. It does so by incrementally moving and updating the centre used for deciding which data should be assigned to a cluster by calculating the mean position from all data points. Further, selecting from within an area around the new mean every move. It takes no parameter to decide the number of resulting clusters, Mean Shift uses *nonparametric kernel density estimates*, to establish an estimation of density it requires only one input parameter, *bandwidth* (Carreira-Perpinan, 2015). It works well with non-spherical shapes (Comaniciu, 2002). In this research it provides a second look at the dataset.

3.3.4 Supervised methods

This section discusses steps and methods takes in conducting supervised learning. Such as generalization and hyperparameter settings, and specific learning algorithms.

Generalization

The data was split into training and testing sets by using functionality available through *Scikit-learn* (Varoquaux et al, 2017). The learning algorithms are trained on one set and evaluated on a selection set aside for testing. For evaluating capability of the model to generalize, *cross-validation* was employed as a method (Sarkar, 2018, p. 289).

Hyperparameters

Prior to training and testing the supervised models, the hyperparameters were set by using the modules *GridSearchCV* and *RandomSearchCV* from *Scikit-learn* (Varoquaux et al, 2017), selecting the model giving the most optimal results. These modules depend on cross-validation to loop through the dataset changing out the testing and training samples by dividing the data at a determined threshold, such as five-fold and ten-fold cross valiation (Claesen, 2015).

Regularization method is set during tuning the models for an optimal combination of hyperparameter settings, determining type of regularization and strength of constriction (Varoquaux, et al, 2017).

Learning rate is set in the *Multi-layered perception* to establish the step size of the *stochastic gradient descent* in the optimizer determined by tuning the model (Varoquaux et al, 2017).

Two regression models are used to predict and classify an outcome, as well as two supplementary methods are used to predict the *survival year* of an individual by a set of explanatory features. The *Linear regression* is used for predicting exact survival year of a prosthetic device, while the rest are utilized for classifying a *survival in year* as grouped outcome classes. Following are the employed models.

Linear regression

Linear regression outlines the relationship between the features passed in as *dependent* (target) and *independent* (explanatory) features. The approach draws a straight line assuming a linear relationship between variables in attempt to minimize the error in forecasting an effect (Sarkar, 2018, p. 315). *Linear regression was* used for predicting a precise outcome prior to primary surgery.

Logistic regression

Logistic regression classifies data within defined categories and differ in type of dependent variable and is used more commonly to determine a bi- or multinomial discrete outcomes (Sarkar, 2018, p. 315). It is a linear model despite the difference in detecting a discrete target, using what is commonly referred to as the *sigmoid function* to *linear regression* to reduce the result to one of a several categories (Sarker, 2018, p. 262). *Logistic regression* was employed

as a baseline method to evaluate classification performance across various learning algorithms that are described below.

Random forest classifier

Random forest classifier is a learning algorithm processesing by assembling multiple decision trees on sub-groups of the data (Varoquaux et al, 2017). By working with multiple trees in parallel the algorithm tries to improve the performance. The collective approach of the decision trees gives an interesting effect as some tress make a poor decision, their collective effort can still move the result in the right direction (Sarkar, 2018, p. 283-284). Random forest classifier was added as an alternative perspective to compare against the other methods.

Multi-layer perceptron

Multi-layer perceptron (MLP) is feedforward neural network, functioning quite differently than the algorithms mentioned above, as it is built on the use of perceptron/neurons as individual processing units in each of the separate layers. *MLP* consists of minimum three layers, one for input, at least on hidden layer for computation, and one output layer for reducing the result to one of the possible outcomes (Sarkar, 2018, p 32). *MLP* was employed as another alternative for comparison, this thesis only used simple *MLP* with three layers.

3.3.5 Evaluation metrics

This section introduces the evaluation metrics employed to evaluate the results in this thesis. There are three separate types, internal cluster validation metrics, and metrics for regression and classification tasks.

Internal cluster validation

Internal validation metrics were used to assess the similarity within and variety between clusters. No proper *ground truth* was known for how to segment data points into groups in any meaningful sense relating to a known outcome.

> ***Silhouette Coefficient:*** This metric is an internal validation metric used when no ground truth labels are present, and attempts to combine the two characteristics, compact and clearly separated, to capture the expected behaviour of good clustering. The metric does so by assessing how similar data points are in relation to others in its own group and how dissimilar they are to those belonging to another group (Sarkar et al, 2018, pp. 280). The result of calculating the coefficient is between 0 and 1, where a higher score means a better result (Rousseeuw, 1987).

***Calinski-Harabasz Index**:* This metric is calculated by considering the ratio of the between cluster dispersion means, i.e. a calculated average ratio of how widely spread the groups are, and how dispersed the data points are within the different clusters (Calinski and Harabasz, 1974). *The Calinski-Harabasz Index* returns a result that is not limited to a number between 0 and 1, and the higher the score the better the result is considered. Similarly used to evaluate without a known *ground truth* (Sarkar et al, 2018, pp. 281).

Regression metrics

Result from *linear regression* were performed by several metrics outlined below.

Coefficient of determinations (R^2): R^2-score measures the likelihood of future predictions being appropriate (Sarkar et al, 2018, p. 281). The best score that can be achieved is 1 and can returns a negative score on low chance of correct prediction. It is used to evaluate *regression* models and test how well the independent features explain the occurrence of the dependent feature (Sarkar et al, 2018, p. 281).

Mean Absolute Error (MAE): *MAE* measures the absolute deviation between a ground truth value known in advance and a predicted value. The metric gives a clear view of how sizable the deviation is overall, illustrated on the same measurement scale as the output (Geron, 2017, p. 39). For instance, if the output is understood in the context of years passed, then the *MAE* score will inform on have many years the predicted value deviated from the known truth by calculation the average from the sum of total errors.

Mean Square Error (MSE): *MSE* performs the evaluation by finding how much a predicted outcome deviates from the true value, considering the average square of the measured deviation. The lower the values returned by calculating the *MSE,* the better the model performs with less errors (Sarkar et al, 2018, p. 282).

Root Mean Square Error (RMSE): *RMSE* is a modification of *Mean Square Error* (MSE), additionally finding the *root* of *MSE*. This metric can also be used to evaluate the performance of a *regression* model, especially the distribution of errors. It is similar to *MAE* in returned value, as it gives a result in the same measurement scale as the predicted value (Geron, 2017, p. 37).

Classification metrics

For assessing the bi- and multinomial classification the accuracy is used to test generalization through cross validation with the accuracy metric. Additionality, the *receiver operator characteristic* provides a second look at performance (Sarkar, 2018, p. 276)**.** They are described below.

> ***Accuracy***: The accuracy score returns an assessment of the overall proportion of correct predictions (Geron, 2017, p. 83). The accuracy score is available through *Scikit-learn* for both bi- and multinomial modelling, giving a consistent evaluation using an identical metric on both classification tasks.
>
> ***Receiver operator characteristic***: The method provides a solution for interpreting the result of classification working well for modelling bi- and multinomial classifiers (Varoquaux, et al, 2017). The curve is created by assembling the *confusion matrix* from the true-positive rate and false-positive rate of performance and plotting the portion true-positive versus false-positive by iterating through the ground truth and the predicated results (Sarkar et al., 2018, p. 276). The *area under the curve*-score provides a further assessment of performance and is measured between 0 and 1 with 0.5 being considered random guessing (Geron, 2017, p. 92). The Receiver operator curve is a common method for evaluating medical diagnostic tests (Hajian-Tilaki, 2013).

4. Technologies and data

This chapter contains a description of the dataset and technologies used to pre-process, explore the contents and training the learning algorithms to perform the regression and classification tasks.

4.1 Technologies

This section presents the different technological tools used to perform the activities in this research.

Python Programming language: Python is a general-purpose programming language, applicable to many domains and supports functional, procedural, and object-oriented programming. It is widely used and has many supporting libraries with tools for data mining and machine learning (Rosssum, 2009). The was used alongside the Anaconda framework.

Anaconda: Anaconda is a free downloadable platform for data science mainly constructed for scientific purposes (*Anaconda Software Distribution*). It allows for setting up a virtual environment with Jupyter Notebook preinstalled and has access to most common libraries used in importing and processing data, as well as access to libraries for visualizing data and machine learning tools.

Jupyter Notebook: Jupyter Notebook was used for programming during the study, it is an application that allows for running Python code live in the browser and is primarily used for cleaning and transforming data, carry out machine learning and statistical modelling tasks, and visualizing data for exploration (Kluyver, 2016). The notebook allows for segmenting code into individual blocks and serves as an effective tool for working through a more adoptable approach.

Scikit-learn: Scikit-learn is a library for data mining, and data analysis in Python, it is an open source solution and available for free (Varoquaux, 2017). Scikit-learn provided the required machine learning tools for exploring the dataset and training prediction models.

Miscellaneous libraries: For working with the data a few noteworthy Python libraries were used, most of all *Pandas*, *NumPy* and *Matplotlib* from the *SciPy* environment. (Virtanen, 2019).

4.2 Dataset description

The dataset is an excerpt of arthroplasty surgery records from the Norwegian National Register for Hip Joint Replacements. In general, the dataset has a minor selection of details on the patients and surgery, reasons for requiring the procedure, and details on materials used in the implanted prosthetic device. The dataset has records collected between the years 1987 and 2018. The selection of patient records and important variables were chosen through discussions with the Orthopaedic clinic at Haukeland University hospital.

As stated above this data is an excerpt and do not contain all the data from the registry, rather it is based on a selection of product types. The types are listed below and is organized around three main products and their combination with a selection of less represented product types.

1. *Cases with Spectron cemented stem combined with cups:*
    1) Reflection cemented HXLPE
    2) Reflection uncemented
    3) Opera
    4) Elite

2. *Cases with Reflection cemented UHMWPE cup combined with stems:*
    1) Spectron cemented stem
    2) ITH stem
    3) Bio-fit cemented stem
    4) Corail stem
    5) Taperloc stem
    6) Hactiv stem

3. *Gold standard cases with Charnley stem combined with cups:*
    1) Charnley

The third selection of cases consists of one type of prosthetic device with a good track record serving as a *gold standard group* with no combinations occurring across product types. Surgery records from the first two groups provide a more varied landscape of different combination to explore.

Column overview

The content of the dataset can be further explained through belonging to different sub-domains:

1. *Patient details*

Organized within the domain '*Patient*' details are the variables relating specifically to the patient, i.e. age, gender, and health status.

2. *Prosthetic device details*

Under the hyponym '*Prosthetic device'* details are the variables directly related to the device, i.e. the materials used in the acetabular cup, femur stem, and the caput, as well as use of polyethylene and size of the prosthetic caput.

3. *Primary surgery reason details*

Accumulated within the domain '*Primary surgery reason*' details are the variables related specifically to reasons for requiring the primary surgery, i.e. the first insertion of a prosthetic device.

4. *Revision surgery reason details*

Organized within the domain '*Revision surgery reason*' details are the variables related to why revision surgery was necessary, such as separate adverse events. These details are only available in records with revision.

A full list of all variables considered throughout exploration in this study is in *Appendix C*, and the different variables are explained more in detail.

5. Data Exploration

This chapter details an exploration phase by invoking clustering as an unsupervised method to examine similarities within and between groups in the dataset, and by describing and visualizing the resulting formations. The primary focus is on locating and exploring possible features for engineering outcomes, and potential explanatory features within the dataset. The purpose is to better understand the predictive powers of the variables and how to further appropriate it for modelling.

5.1 Approach

The approach taken in exploring was performed in the manner detailed here. The following two steps are first performed.

1. The data is checked for missing values.
2. The features to include in clustering models are selected.

The data is separated into three different sets of records, those with a device survival length at < 5, < 10, and < 15 years. For each of the sets of records the following step are performed (Figure 5.1):

1. The number of clusters is determined by running *the Elbow method* **(**Kodinariya, 2013**)** using *K-means* to locate a range of possible K-values to explore the dataset.

2. Clustering is performed between a range of possible K-values, after each clustering the *Silhouette Coefficient* and *Calinski-Harabazs index*, as well as details on records in individual clusters, are summarized in tables, *Appendix D-F*.

3. The Mean shift hyperparameter *bandwidth* is estimated by using *Silhouette coefficient*, after clustering is done the details on records in the set of produced clusters are listed in tables, *Appendix J and K*.

4. Variance between clusters is examined, *Cramér's V* correlation coefficient (Akoglu, 2018) are employed to assess explanatory power in independent features toward possible outcome features in the overall populations and in a selection of individual clusters.

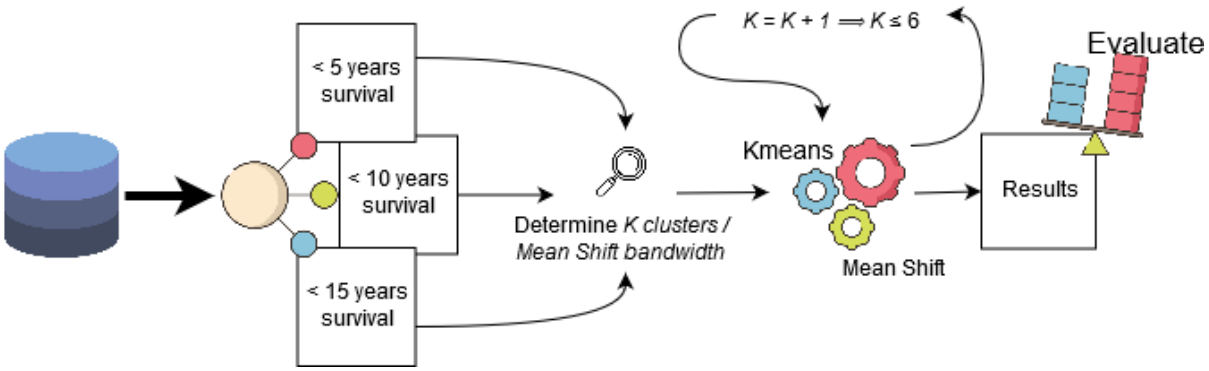

Figure 5.1: Step performed with clusters detailing initial segmentation of records, determining hyperparameters, and processing the data.

5.2 Distribution of values

An inspection of missing values in the dataset with all cases, revision and non-revisions surgeries, show that most columns are without *Nan* values, in total only three columns have a minority registered values available. For a second perspective on what is in the data (Figure 5.2) included the distribution of values that are *positive*, *Nan* and 0.

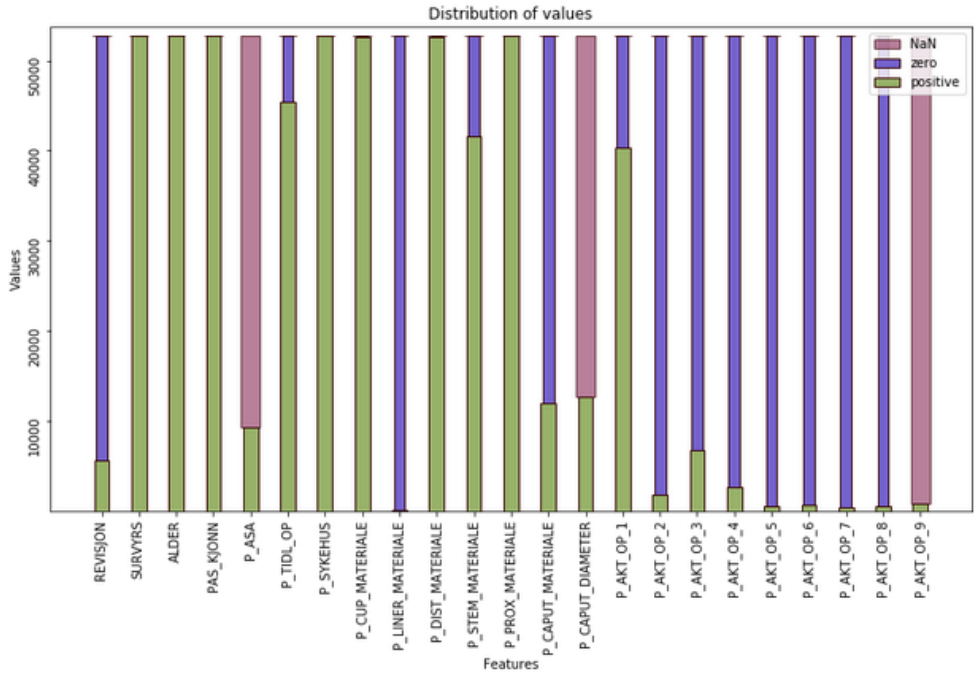

Figure 5.2: Distribution missing (marked red), 0 (marked blue) and present values (marked green).

Looking toward the columns denoting the use of materials in separate parts of the prosthetic device there is more available positive values than there is in reasons for requiring primary

surgery. Variables on reasons for requiring primary surgery show that in the majority of records the reason for hip replacement was *coxarthrosis* and only a minor selection originates from another cause, reflecting the majority *coxarthrosis* in the overall registry (Furnes, 2019)

5.3 Features selection

Features for clustering were selected through a process of elimination where the features detailing knowledge inaccessible prior to primary surgery were removed. Thereby excluding features such as surgery length and details on why cases required *revision* surgery. The selected features are listed below, and primarily concern two separate domains, *case specific* and *prosthesis device specific* details. The details on the device include materials used in the procedure and size of the *caput*.

For handling categorical variables without a contextual relationship between values, all features with a multinomial finite scale of values was pre-processed by *one-hot encoding* creating a binary matrix representation of the categorical variables.

Standardizing was than performed to reduce the effect of a variety in types of measurements within the data as there are multiple dimensions with different natures. Among them age and caput size, one in years and the other is centimetres.

Patient specific variables:

ALDER: Describes the age of an individual at the time of surgery and is represented by a continuous value measured in years.
PAS_KJONN: Feature describes the gender of an individual and is represented by a binary value.
P_ASA: Describes the health status of an individual and is represented by a categorical value between 0 and 5, and 9, with 9 representing unknown and 0 being unregistered/left blank.
P_TIDL_OP: Describes if a patient has had surgery in the hip outside of hip replacement and is represented by a binary value.
P_SYKEHUS: Describes at which hospital the surgery took place and is represented by a categorical variable representing different hospitals.
REVISION: Describes if a patient had revision surgery and is represented by a binary value.
SURVYRS: Describes the length of survival for an individual case/device and is represented by a continuous value measured in years.

Prosthesis device specific variables:

Each categorical variable can be between 0 and 12 and the variable describing prosthetic caput size is stored as a continuous variable. In both categorical and continuous variables, the value 0 represents an empty column.

P_CUP_MATERIALE: Describes the materials used in the cup between the caput and stem and is represented by a value between 0 and 12.
P_LINER_MATERIALE: Describes the type of polyethylene liner between cup and caput and is represented by a value between 0 and 12.
P_STEM_MATERIALE*:* Describes the value used in the stem of a device and is represented by a value between 0 and 12.
P_PROX_MATERIALE*:* Describes the type of polyethylene liner between the femoral-head and cup and is represented by a value between 0 and 12.
P_DIST_MATERIALE*:* Describes secondary materials used in the stem and is represented by a value between 0 and 12.
P_CAPUT_MATERIALE*:* Describes materials used in the head of the component and is represented by a value between 0 and 12.
P_CAPUT_DIAMETER*:* Describes the size of the head of the component and is represented by a continuous value.

A complete list of materials and their coding is in *Appendix B.*

5.4 Clustering with K-means

This section is a walkthrough of results from model-based clustering with K-means from *Scikit-learn*. Details on resulting clusters are documented in *Appendix D-F*.

Produced clusters are as well compared to the original groups based on main product type briefly presented in Chapter 4, although these do not represent a *ground truth* in relation to any known occurrence of an outcome.

In this section the terms «groups/original groups» are used for referring to the organization of device types presented in Chapter 4 and «gold cases» for group 3 with only one product type overall (Charnely). While «cluster x» is used frequently and distinguishes between resulting clusters.

5.4.1. Determining number of clusters

The *Elbow method* was used for determining an appropriate number of clusters (Kodinariya, 2013). This was done starting at $K=2$ and ending at $K=10$ clusters for the three sets (< 5, < 10, < 15), with the *original* formatting of the data and with data *standardized* using *Scikit-learns* module *StandardScaler* (Buitinck, 2013).

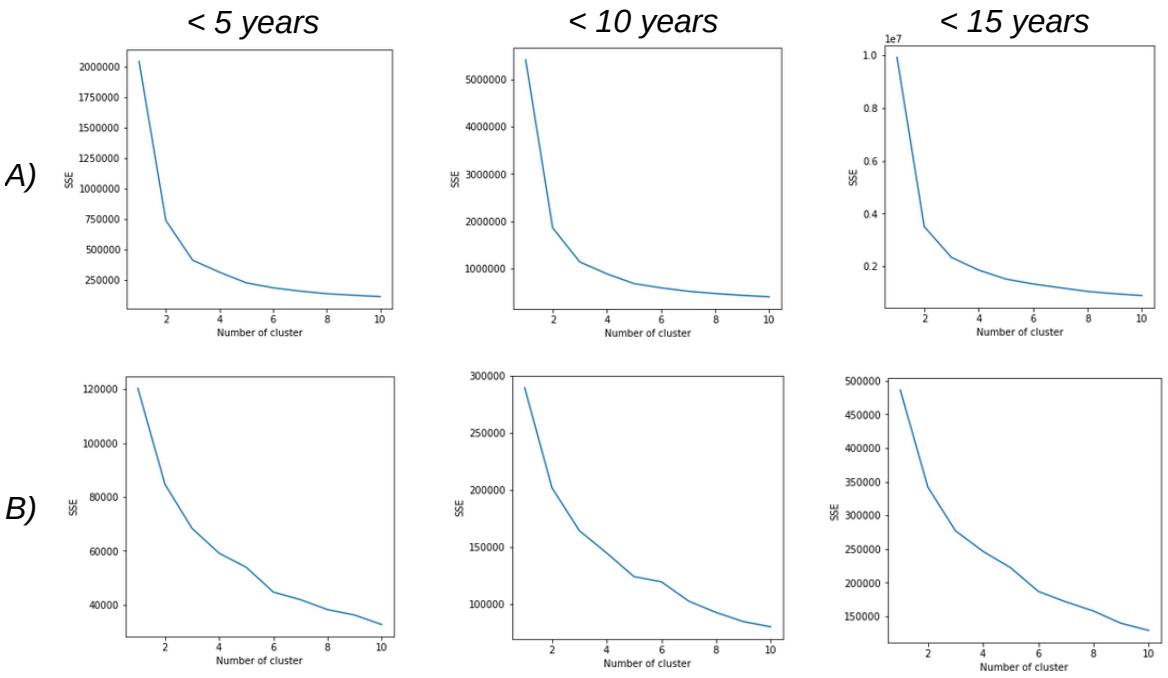

Figure 5.3: Elbow method with regular formatted data vs standardized data, in row A) the data is not changed, in row B) the data is standardized.

The results show a sharp angle with data in the original formatting that is more distinguishable at three clusters and persistent across the < 5, < 10 and < 15 years. While the standardized data show a more ambiguous result, the angle appears at five and six clusters for data < 5 and < 10 years survival. The results from < 15 years indicate a change in distribution of records occurring at six clusters, however there are less distinct changes in convergence across the spectrum than in the smaller selections of data (Figure 5.3).

Taking into consideration the minimal variety in product types in the dataset with only a handful of prosthesis products in the overall population, the area to search were set to $K=2$ and $K=6$.

5.4.2 Survival below five years

The overall results from cluster $K=2$ to $K=6$ with a selection of records with survival at < 5 years prior to requiring revision surgery is in Appendix D. Here are the main findings.

Overall population: There are 9257 records, the total number of unique materials is *seven*, with a majority steel and chrome/aluminium stems. The number per group is 325 Reflection UHMWPE, cup combined with a range of stems, 1526 Spectron stems combined with a range of cups, and 7403 Charnely type from *gold records*, with mean length of device survival before revision at 1.7, 2.1, and 2.4 years respectively.

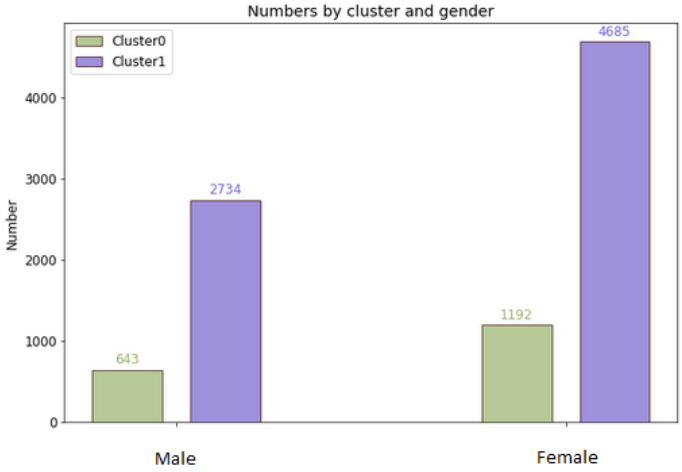

Figure 5.4: Distribution of records in clusters according to gender with data < 5 years.

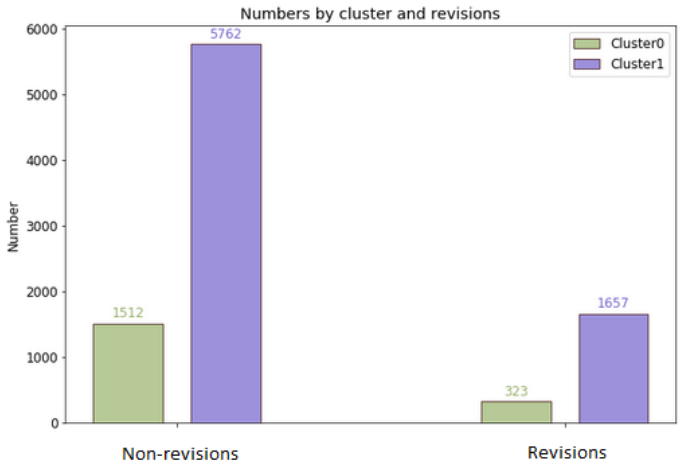

Figure 5.5: Distribution of records in clusters according to revision surgery with data < 5 years.

.

K* = *2: The two smaller selections of records are made into one cluster, and the majority of *gold records* is alone in cluster 1. There is no unique division between clusters and gender, the characteristic in the overall population of a majority female is represented equally in both clusters (Figure 5.4).

The revision rate is similar in both clusters, at 17.6% and 22,3% (Figure 5.5). In survival years there is a slight deviation, with a mean survival in cluster 0 occurring 4 months earlier than cluster 1. In materials the clusters appear to be separated by one larger difference, cluster 1 include almost all records with the same use of *steel* contrary to a wider variety observed in cluster 0.

K* = *3: Results with *three* cluster show the larger group of *gold records* is still together in one cluster, while clusters 1 and 2 include a large variety. These clusters have a larger difference in records with Spectron stem and type of polyethylene.

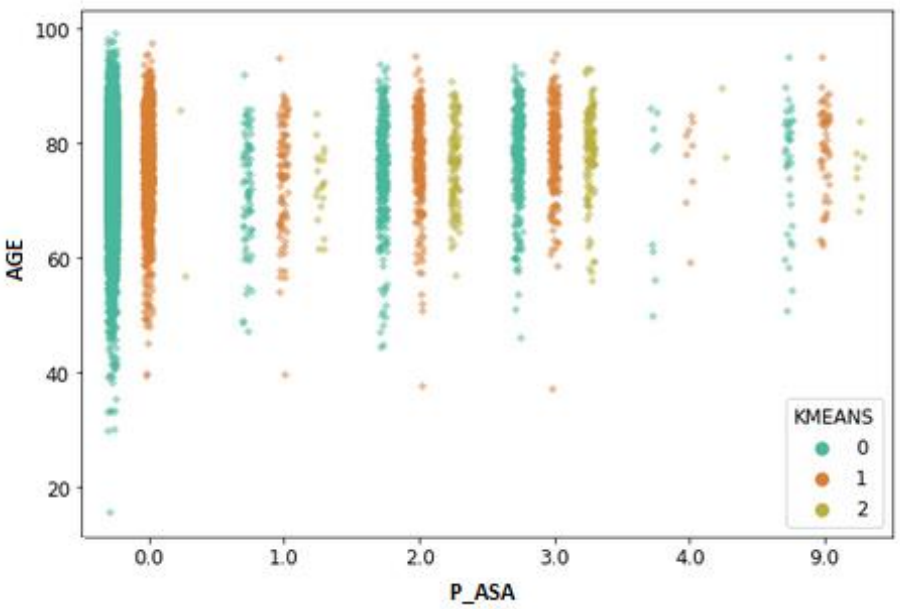

Figure 5.6: Distribution of records in clusters according to health status by The American Society of Anaesthesiologists (ASA) indicator with data < 5 years.

There is no significant distribution of cases among *genders*, though there is a clear distinction between cluster 2 and the others in number of records with registered *ASA-class* (Figure 5.6), as it has practically no missing values. On revisions the rate is 22%, 17%, and 15% and length of survival is 2.4, 2.1 and 1.4 years respectively.

K = 4: There is barely any new variation in use of materials between the clusters, the new change occurs mostly within a selection of similar steel cases from the cluster of *gold records*. There is no distinct difference in gender or health status among them.

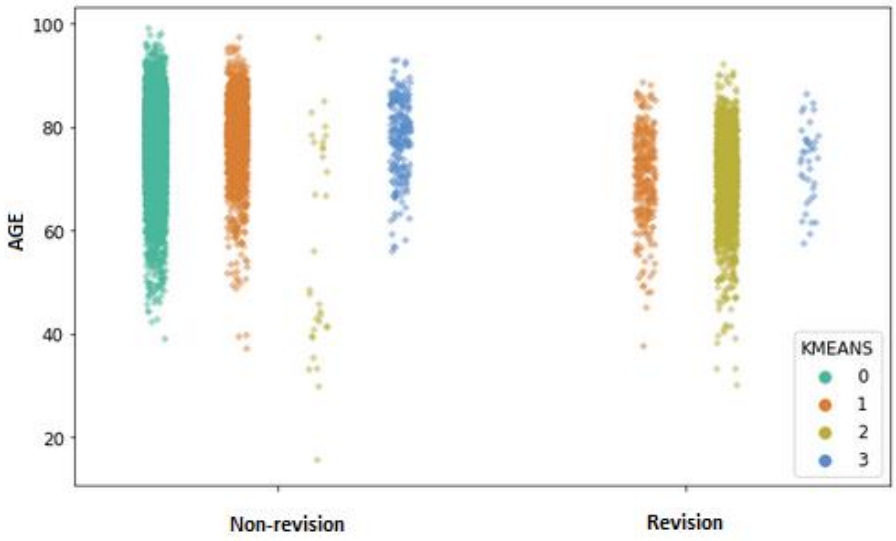

Figure 5.7: Distribution of records in clusters according to revision surgery with data < 5 years (4 cluster in total).

Cluster 0 has no revision, and consists of 5715 *gold cases*, and cluster 2 is comprised of 98% revision surgeries and has most of the remaining out of all 7400 *gold records* < 5 years device survival. The mean survival year before requiring revision in cluster 3 is 2,4 years with 98% revision surgeries (Figure 5.7), Clusters 0 and 1 have survival at a persistent 1,4 and 2.1 years.

K* = *5: Some small changes, but not much alteration in distribution is observed. Cluster 4 contains 13 records with Reflection Uncemented cup, with 8 revisions and 84% male population and a mean survival at 2.4 years. The clusters with 1,4 and 2,1 years are persistent with only minor alteration (Figure 5.8).

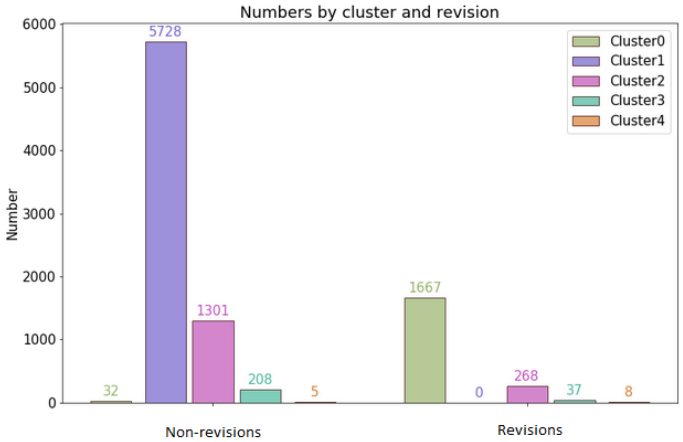

Figure 5.8: Distribution of records in clusters according to revision surgery with data < 5 years (5 cluster in total).

K* = *6: There are some larger changes in how the records were distributed related to gender with six clusters, clusters 2 and 5 are established without revisions and separated by gender (Figure 5.9). Cluster 5 with all female records and cluster 2 with 99.8% males, however, there are no other distinct changes in materials between these cluster and the others.

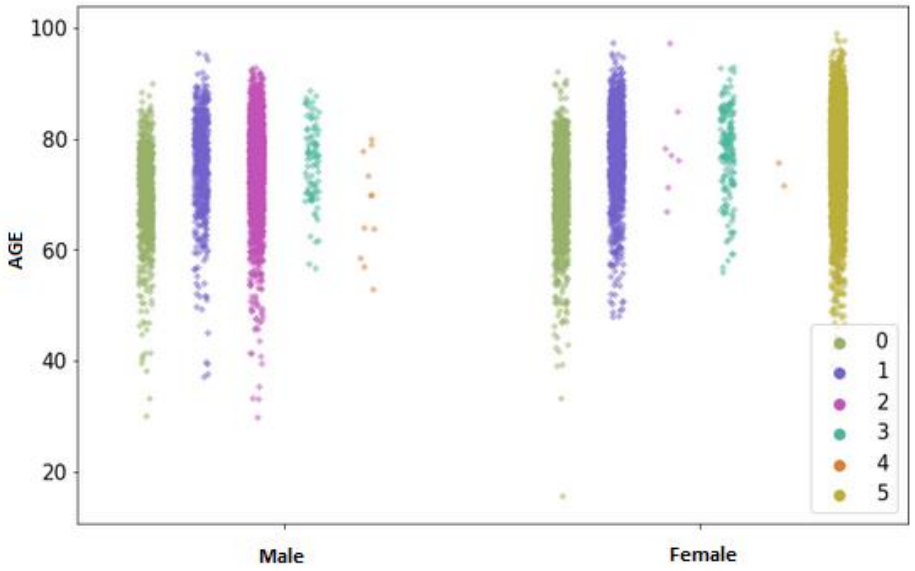

Figure 5.9: Distribution of records in clusters according to gender with six clusters and data < 5 years (six clusters in total).

Brief summary for < 5 years: There is some persistency throughout $K=2$ and $K=6$ in mean survival between clusters, and a distinct similarity in materials within clusters. Reoccurring mean length of survival outcomes are *1.4*, *2.1* and *2.4* in years, as well as repeating cluster with all non-revisions. Looking at reasons for requiring primary surgery the majority resulted from *coxarthrosis* and *fractures* (Figure 5.10), and a dominant *aseptic loosening of the stem* as main adverse events leading to revision surgery.

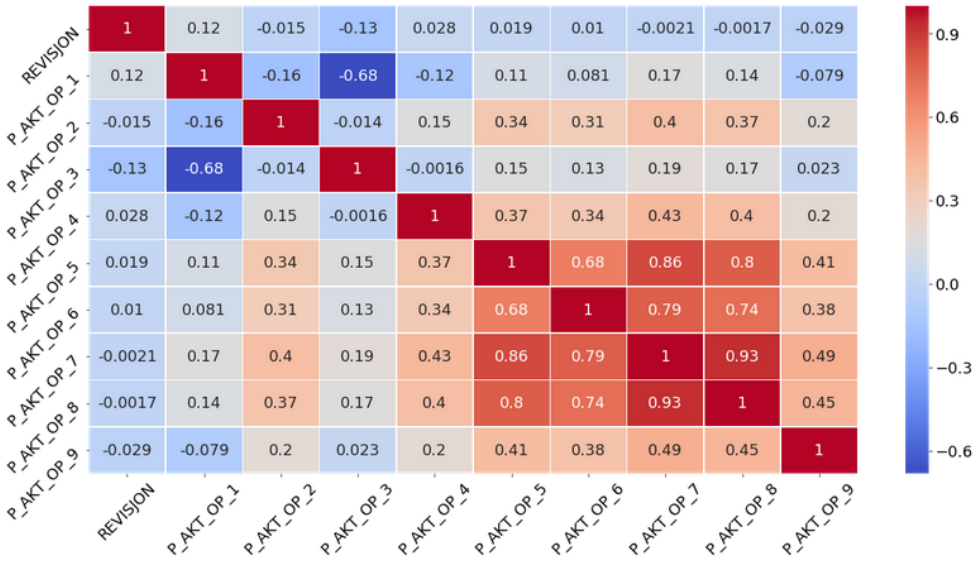

Figure 5.10: Cramér's V correlation coefficient for categorical primary surgery reasons and revision indicator in all records with survival < 5 years.

5.4.3 Survival below ten years

The overall results from cluster $K=2$ to $K=6$ with a selection of cases with survival at < 10 years prior to requiring revision surgery is in *Appendix E*. Here are the main findings.

Overall population: There are in total 22 272 records < 10 years device survival, with mean survival of 4.2, 5.7, and 4.4 years organized by product types in the following order. The amount per group is 1302 Reflection UHMWPE cup combined with a range of stems, 3967 Spectron stems with a range of cups, and 17003 Charnely type of *gold records*.

K = 2: Minor change is observed in the way data is grouped in comparison to the data with $<$ 5 years device survival. The *gold records* are together in cluster 0, while the more disperse combination of products is in cluster 1. The rate of revision is 5.6 years for cluster 0 and 4.4 years for cluster 1. There is no significant difference in revision rates with 16.1% and 16.5% percent of records having had re-surgery before ten years.

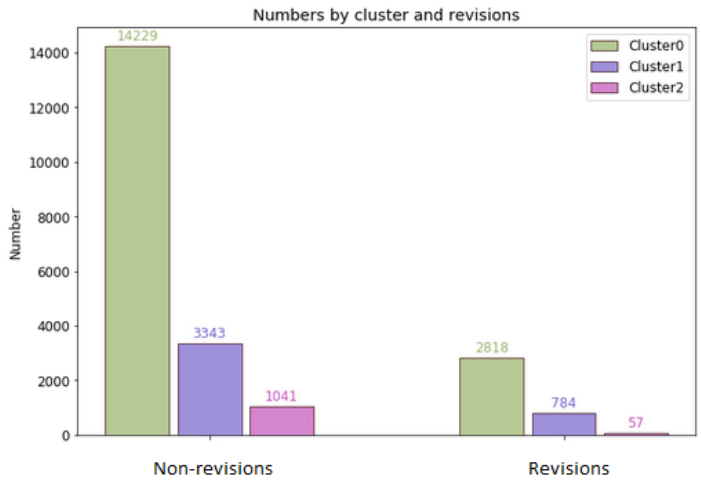

Figure 5.11: Distribution of records in clusters according to revision surgery with records < 10 years.

K* = *3: Results indicate a similar organization of records as seen < 5 years. Records with the majority of Spectron alumina stem *and a* sharing a similarity in type of polyethylene is moved to the new cluster 2.

The distribution of records related to gender is 65%, 66%, and 67% female, reflecting the overall population. Cluster 2 with 99% *ASA-status* present among its records is similar to the one < 5 years. (Figure 5.12). The rate of revision is 16%, 15%, and 5% per cluster, with cluster 2 having a significantly earlier revision at 3.6 years (Figure 5.11). Cluster 0 and 1 have 4.4 and 5.7 years survival, respectively.

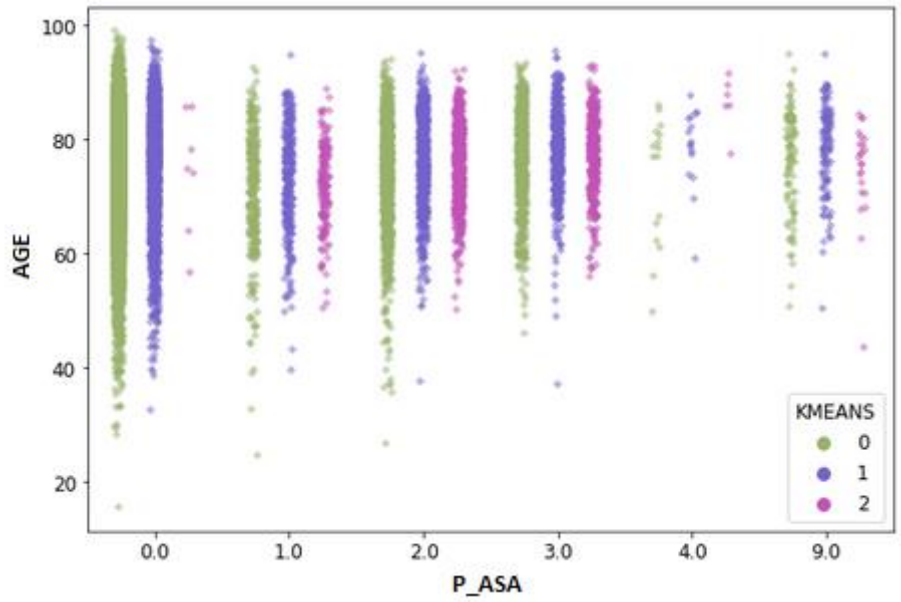

Figure 5.12: Distribution of records in clusters according to health status by The American Society of Anaesthesiologists (ASA) indicator with data < 10 years (3 clusters in total).

K* = *4: The change seen by increasing the number of clusters is mainly due to the larger structure of *gold records* with steel prosthetic as it splits into clusters 1 and 2. While the previous clusters with a wider variety in product types have no clear alteration. Gender and ASA-class distribution express no difference between previous results.

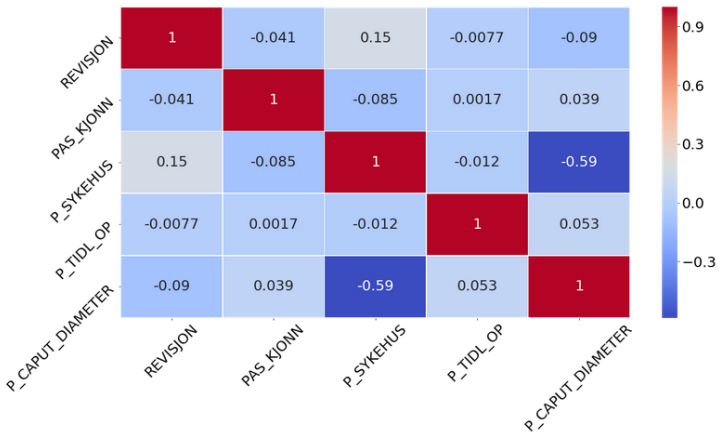

Figure 5.13: Cramér's V Correlation coefficient for continuous and categorical variables in K = 4, cluster 4 (<10 years survival length).

There is a split in revision and non-revision records among the clusters lacking variety in materials. Cluster 1 has no revision and cluster 2 has approximately 99% revisions. The mean length of survival among revision cases is persistent at 5.7, 4.4 and 3.6 years, excluding the cluster with no revision surgery. The patient specific variables showed no clear significant correlation toward the occurrence of revision (Figure 5.13).

K* = *5: Changes are minor, a new cluster appears with 44 records from the persistent selection with a mean survival of 3.6 years, now with 5.1 year in survival length and a difference in use of cup type and a larger number of undocumented cup material. There is no significant difference in the distribution of gender in the overall population except for the new cluster 2, with has approximately 50% split between male and female.

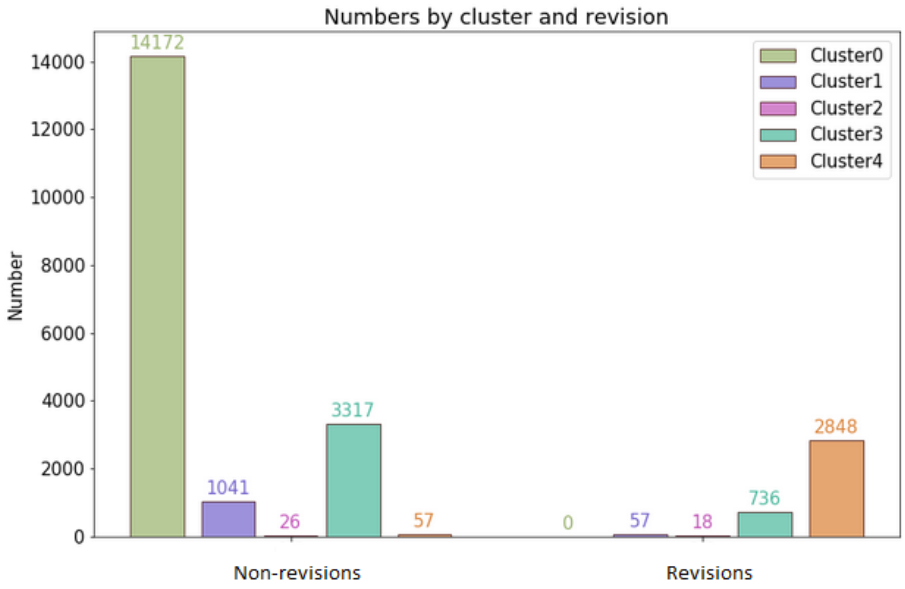

Figure 5.14: Distribution of records in clusters according to revision surgery with records < 10 years and five clusters.

Revision rates are 0%, 5%, 40%, 18%, and 99%, there is persistent distribution of records among clusters seen previously, as well as the new cluster 1. Though this cluster is only a minor selection of records (Figure 5.14**).**

K = 6, has some significant changes to the cluster with no revision mentioned above, as it forms cluster 0 and 4. The difference is largely in gender, with cluster 0 having only females and cluster 4 with 99.7% male records, and a minor difference among cases in use of alumina in the *caput material* (Figure 5.15). The rest of previously seen clusters are rather persistent, with only small changes in distribution.

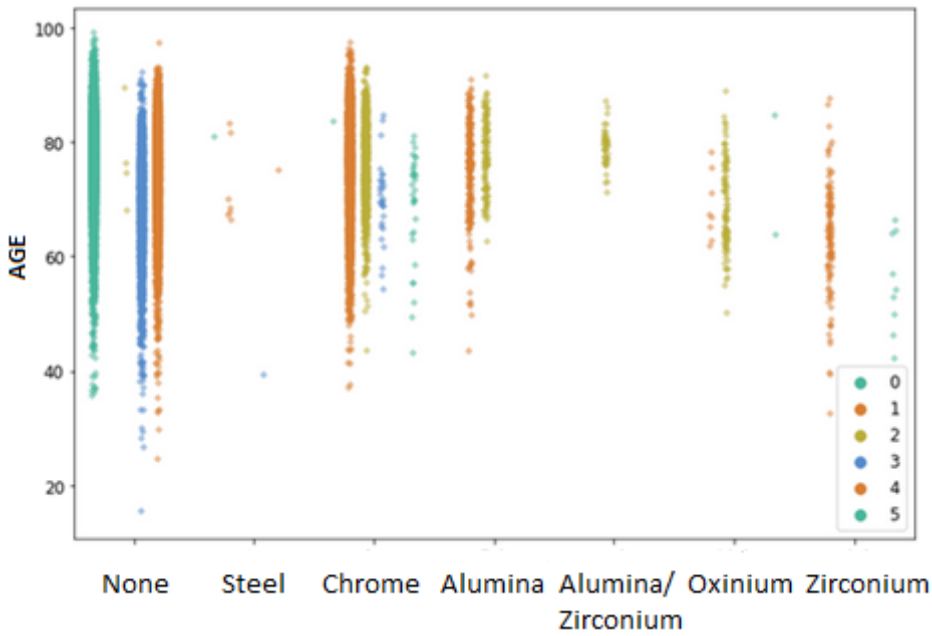

Figure 5.15: Distribution of records in clusters according to caput material with records < 10 years and six clusters.

The revision rates are at 0%, 18.1%, 5.1%, 99.6%, 0% and 40% ordered from cluster one to six. The observed clusters have a large similarity toward previous results, with persistency in cluster 1, 2, 3 and 5, with a survival in years at 5.7, 3.6, 4.4, and 5.1 years, excluding clusters 0 and 4 with no records with revision.

Brief summary for < 10 years: Persistence in distribution of records in clusters is observed in < 10 years, as it was in data < 5 years. There is consistency in how clusters are differentiated by prosthesis product types and a largely common us of polyethylene type within larger clusters. In mean survival years and revision rates there is a frequent reoccurrence of outcomes, in year *3.6, 4.4, 5.7,* and a persistent two formations with no revision surgeries. Examining *reasons for primary surgery* among clusters show a similar

phenomenon in a majority *coxarthrosis*, as well as a considerable amount of *fractures* and *rheumatoid arthritis*. This is similar to reported statistics (Furnes, 2019) indicating primary reason is *coxarthrosis*, and a decline in *rheumatoid arthritis* in later years.

5.4.4 Survival below fifteen years

The overall results from cluster $K=2$ to $K=6$ with a selection of cases with survival at < 15 years prior to requiring revision surgery is in *Appendix F*. Here are the main findings.

Overall population: In total there are 37406 records < 15 years survival, with mean survival of 5.7, 8.2, and 6 years organized by product types in the following order. The amount per group is 2161 Reflection UHMWPE cup combined with a range of stems, 7828 Spectron stems with a range of cups, and 27420 Charnely type *gold cases*.

K* = *2: has no clear distinct difference between clusters in gender, and a similar distribution of records as observed on data < 5 and < 10 years. The separation is almost completely clear between *gold records* and the smaller groups of variations (*Spectron/Reflection* product types) with only a 0.2% overlap among all three clusters.

***K* = *3*:** There is little change occurring as compared to the results from < 5 years, they are rather identical. The clearest observed difference in materials is the use of *highly cross-linked polyethylene* between clusters 1 and 2 (Figure 5.16). Cluster 0 has majority *gold records*, a similar distribution as seen previously. The revision rates are 12.7%, 14%, and 3.2%, with a mean survival length at 6, 8.2, and 3.6 years, the lowest revision rate also corresponds to the cluster with earliest occurrence of revision.

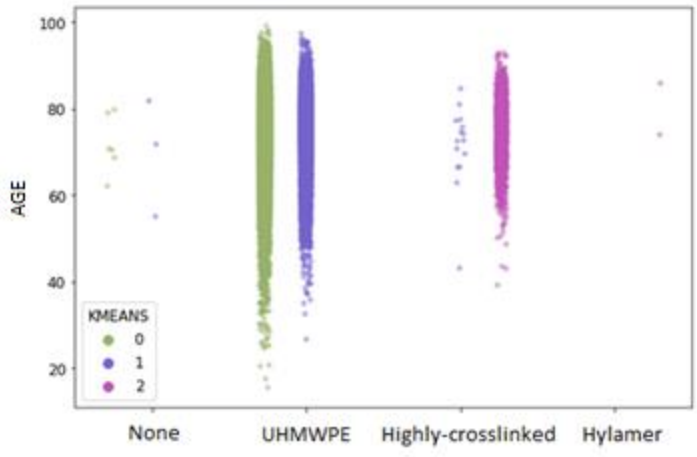

Figure 5.16: Distribution of records in clusters according to polyethylene liner with records < 15 years and three clusters.

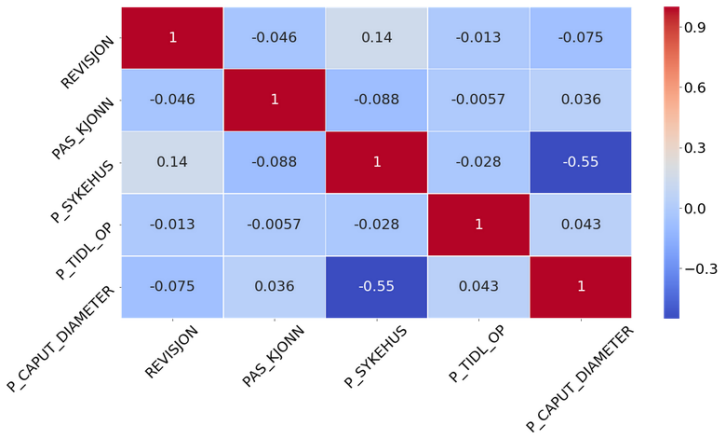

Figure 5.17: Cramér's V Correlation coefficient for K=4, with records from cluster 3 and 8.17 years device survival (data < 15 years survival length).

***K = 4**:* There is a similar distribution as seen prior, the clusters with a wider variety and a smaller number of records remain largely unchanged. While the group of *gold records* divides into two, cluster 0 and 3, distinguished by a clear separation in revision and non-revision. There is persistency in clusters with 3.6 and 8.17 years survival (Figure 5.17), and the new cluster with only 4 non-revision records has a mean survival at 6 years.

K = 5: An interesting change in the larger selection of *gold records* occurs, altering the distribution among revision records seen with $K=4$ clusters. Two clusters are established, one with 3.7% and the other with 23% revision rate. There is larger deviation in mean survival years among records with a positive revision indicator, with 12.7 years in cluster 0 (Figure 5.18) and 4.7 in cluster 3. Furthermore, there is a new smaller cluster originating from group 1 with only 70 records, 35.7% rate of revisions and a mean survival at 7 years, with a clear difference in use of *polyethylene* type. The only small distinction in distribution among gender and patient condition is the new cluster of 70 records with 48.5% male population.

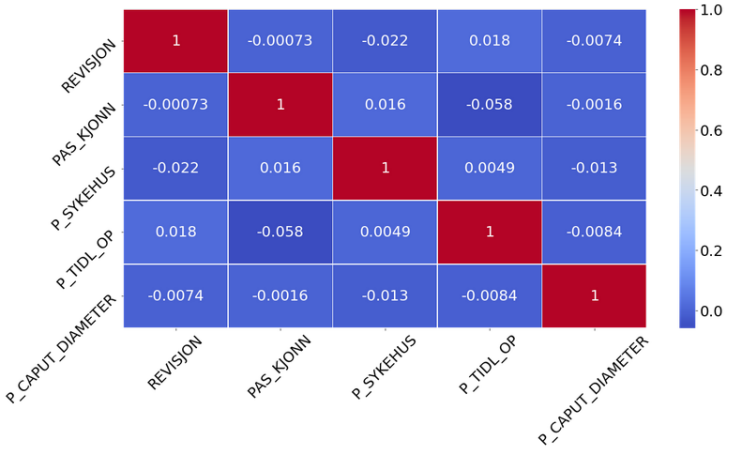

Figure 5.18: Cramér's V Correlation coefficient for K=5, with records from cluster 1 and 12.7 years survival (data < 15 years survival length).

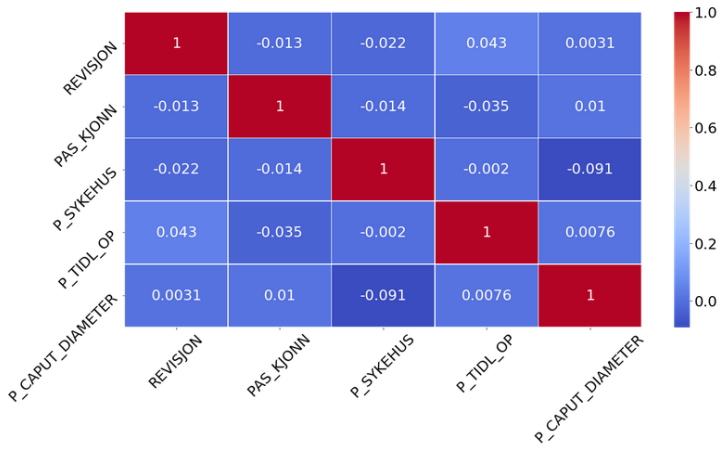

Figure 5.19: Cramér's V Correlation coefficient for K=6, with records from cluster 4 and 6 years device survival (data < 15 years survival length).

K = 6: Increasing number of clusters changes the distribution according to gender among *gold records* without revision, a phenomenon observed in data with < 10 years device survival. Cluster 1 is 99% male records, while cluster 2 is all female, both with no revisions. The revision rates are 23%, 99%, 35% and 3.2% in clusters 0, 3, 4, and 5, while mean survival length before requiring revision surgery is 8.2, 6, 7, and 3.6 years, excluding clusters with no revisions (Figure 5.19).

Brief summary for < 5 years: Survival outcomes in years have some reoccurring similarities in a few general areas, such as reasons for primary (Figure 5.20) and revision surgery. The least lasting cluster is persistent at *3.6* years survival, similarly observed in data < 10 years and < 5 years (1.4 years). Frequent mean device survival lengths are 3.6, 6, 8-8.2 years, and another interesting outcome only seen once was a sizeable selection of revision records at 12.7 years, the longest mean survival observed in any cluster.

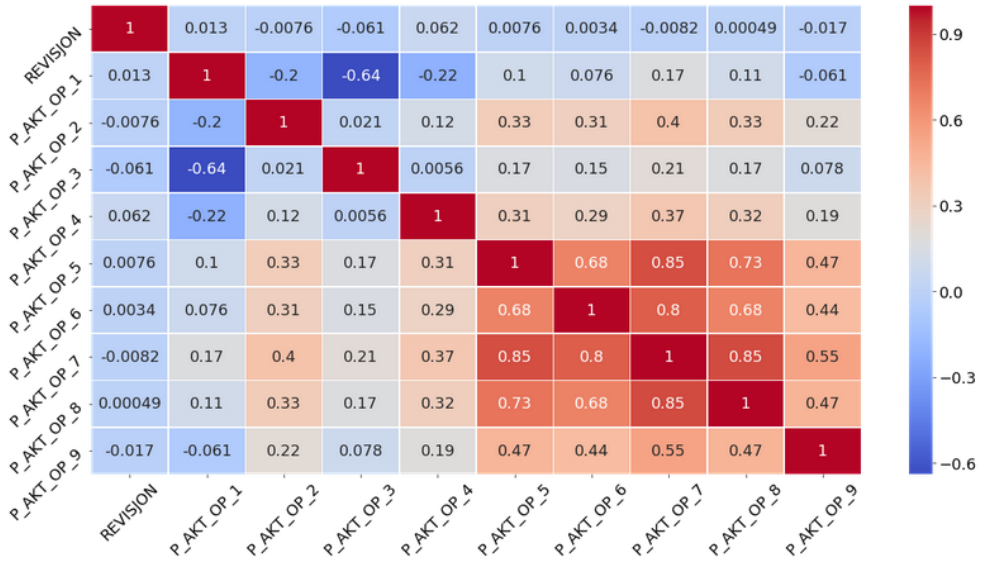

Figure 5.20: Cramér's V correlation coefficient for categorical primary surgery reasons and revision indicator in records with survival < 15 years.

5.5 Clustering with Mean Shift

This method was chosen as a means for comparing results between separate clustering models. Mean Shift is sensitive to any estimation of the *bandwidth* parameter, increasing the *bandwidth* can cause data points to merge at a higher rate, while turning it down gives less convergence between data points (Carreira-Perpinan, 2015).

In this study estimating the hyperparameter was done by recording the *Silhouette Coefficient* to assess cluster structures with different bandwidth-values (Table of results for estimating bandwidth are in *Appendix L*). As a result, the parameters were set to 0.35, 0.275 and 0.25, in the order of < 5 years, < 10 years, and < 15 years.

5.5.1 Survival below five years

The table of results is in *Appendix J*. Starting at a population size of *9257 records*, there is a much similarity in the larger clusters, 1, 2, and 3, to what was seen in *Section 5.2*. While several smaller clusters ranging from 1 to 52 members have been distinguished as their own formations. The difference in materials reflect a separation between steel and chrome/alumina, or alumina-based combinations observed previously, and a reoccurring cluster 3 with 37 revision cases and 1.4 years mean survival before revision. Looking toward gender and registered patient condition there is no distinct difference between resulting clusters. Although most outliers are females, they have none or only a small amount of revision surgeries.

The interesting clusters are two of the smaller ones, 6 and 12. Cluster 6 has 44% revision and the only cluster with the material *Zirconium*. Cluster 12 has only 4 members and distinguishes itself from the rest by the average age of patients at 29 years, although there is no clear difference in use of materials at a lower patient age in this data.

5.5.2 Survival below ten years

The table of results is in *Appendix J*. Increasing the population size (*22 272 records total*) do not bring any significant changes to the distribution in accordance with use of materials. There is a clear separation between the steel and chrome/alumina records, and a minor degree of overlap between alumina and the less represented materials. There is also a persistence seen in cluster 5, with *Zirconium*, now at a revision rate of 59% and an average survival at 6.78 years.

5.5.3 Survival below fifteen years

The table of results is in *Appendix K*. A total of 37406 records with device survival < 15 years has no distinct difference occurring in distribution of cases on grounds of materials used in the device, the same steel and chrome/alumina differentiation is dominant. Among the smaller clusters forming the most interesting is the persistent cluster 5, with a revision rate at 51% and an average survival length prior to requiring revision at 9.33 years. Cluster 5 has the second longest mean survival in years, only surpassed by a one-record cluster, number 14, with 9.81 years.

Brief summary for Mean Shift: Overall, there is a larger similarity towards the original presentation organized by product types and towards results from *K-means*. There is some difference in reasons for revisions between the two larger clusters, with aseptic loosening of the stem as dominant reason in clusters with majority steel records and loosening of the cup in clusters with a majority alumina-based device. The only reoccurring cluster of interest is cluster 6 in data < 5 years, and cluster 5 in data < 10 years and < 15 years survival, with a persistent revision rate at above 50%. It is additionally one of the formations with a longer average survival among revision records. Among smaller clusters there is a repeating phenomenon of all cases having an unknown patient condition (*ASA-class*), however these do not correspond to any higher rate of revision.

5.6 Summary

Silhouette Coefficient has the highest result on two or three clusters, however the indication is not a particularly strong cluster (Table 5.1) (Rousseeuw, 1987). Similar result is returned by the Calinski-Harabasz Index (Table 5.2), regardless, on visual inspection the $K=2$ and $K=3$ did not produce specifically interesting results in relation to outcomes or risk groups.

Table 5.1: Silhouette Coefficients for K <= 6 for data < 5, < 10, < 15 years device survival.

K	< 5 YEARS	< 10 YEARS	< 15 YEARS
K=2	0,473	0,444	0,424
K=3	0,471	0,448	0,434
K=4	0,314	0,347	0,359
K=5	0,321	0,354	0,238
K=6	0,305	0,318	0,329

Table 5.2: Calinski-Harabasz Index for K <= 6 for data < 5, < 10, < 15 years device survival.

K	< 5 YEARS	< 10 YEARS	< 15 YEARS
K=2	3891	9692	15812
K=3	3523	8501	14139
K=4	3123	7437	12147
K=5	3190	7444	11511
K=6	3127	7329	11991

Investigating *reasons for requiring the primary surgery* gave a description of potential features, although the correlation between *revision* and *primary reasons for requiring surgery* were largely insignificant. Primary surgery reasons 1 and 4, have a barely positive coefficient observed across the selections of data; checking individual clusters did not indicate any correlation between revision and other variables. Moving up to all data < 15 years the same reasons have a correlation closer to zero, indicating a neglectable effect on revision in the larger sample of data.

Materials showed to be divided between clusters largely on stem and caput materials, with clusters forming as either large and consisting of a similar blend of well represented materials, or smaller with a variety of less represented materials. Clustering with a combination of materials used in different parts of the prosthetic device did present a range of formations with a variety in mean survival outcomes occurring on a broader range. Exploratory analysis additionally led to three issues in establishing a dataset with a *ground truth* available for modelling:

The undecided: Currently the survival year is not an absolute length of survival until revision unless the revision indicator is positive. Most cases do not have revisions and for those records the listed device survival length is an ongoing process and is not appropriate for a *ground truth* to indicate time from primary surgery to revision surgery.

Those whom left us: The mean age at primary surgery in the dataset is 68 years, in a considerable number of records the patient has passed on before any problems occurred or moved away, and the information of whether a revision surgery happened is unknown.

Similarity in reason: Correlation between causes and revision indicator was low, with only two causes for requiring primary surgery showing a slight positive result. Increasing the feature space in modelling by including all reasons may confuse more than it can help. The reasons for requiring revision surgery do render an opportunity for classifying an expected outcome. However, the majority in our dataset is aseptic loosening and the variety between records is sparse with a majority of one type of prosthetic device present.

Clustering gave a few interesting results inspecting outcome in years of survival between clusters with a difference in product types and their materials, however, there were no clear indication of groups with a specific risk of a certain adverse event leading to re-surgery. The mean survival of devices in years were somewhat persistent and reoccurred across the data separated by the thresholds in survival year. Results indicated a difference in when revisions occurred for separate groups of patients, suggesting using the existent revision indicator in the dataset for predicting revision or no revision may not be informative. As an individual patient seeking to know more about the chances for revision surgery may be confused without additional information on time, place and reason the revision might be required.

5.6.1 Features

Feature engineering

Engineering outcome features was done by establishing two new variables, one multiclass defined as < 5 year as class 0, ≥ 5 and ≤ 10 year for class 1, and > 10 years survival before revision surgery for class 2. In addition, another outcome feature was established by defined records as either ≤ 8 as class 0 and > 8 years device survival for class 1. This was done to add relatable context to the issue of predicting if a revision would occur (Figure 5.21 and Figure 5.22). It also represents the result from the exploratory analysis showing that different clusters varying in materials had a range of results from a few years after surgery to as late as twelve years after primary surgery.

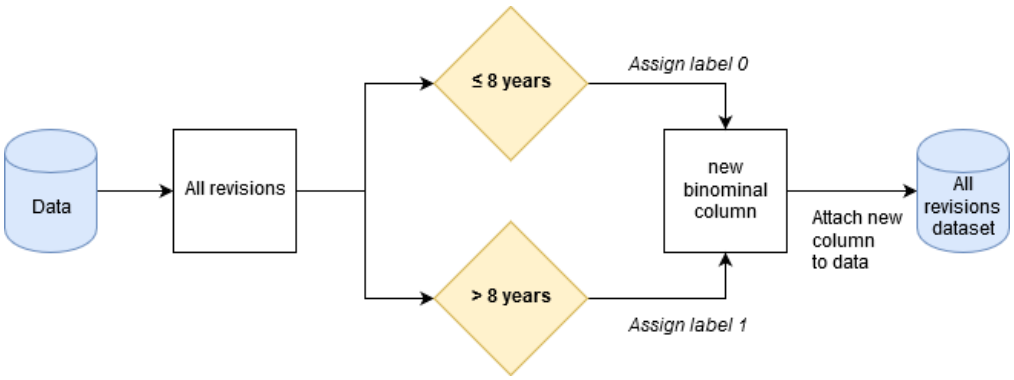

Figure 5.21: Steps taken to add a new column for representing binary class of survival outcomes.

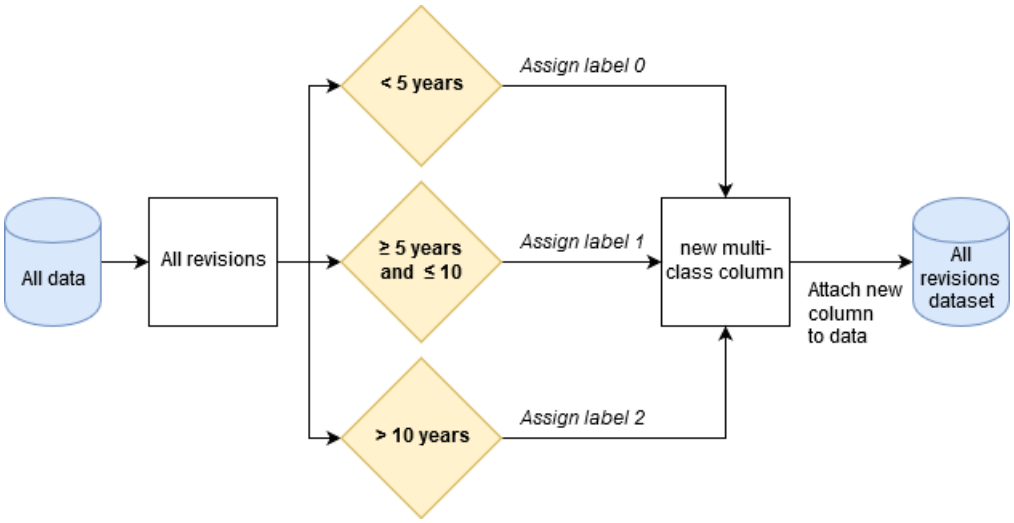

Figure 5.22: Extraction process to add a new column for representing multi-class of survival outcomes.

Another possibility is to include primary operating reasons, although the data is heavily skewed toward one of the reasons (coxarthrosis). All reasons above *P_AKT_OP_4* are hardly present and often occur together, i.e. correlate with each other. To include primary surgery reasons further without adding to many dimensions and to avoid noise the sparsely represented reasons are grouped into one column (Figure 5.23).

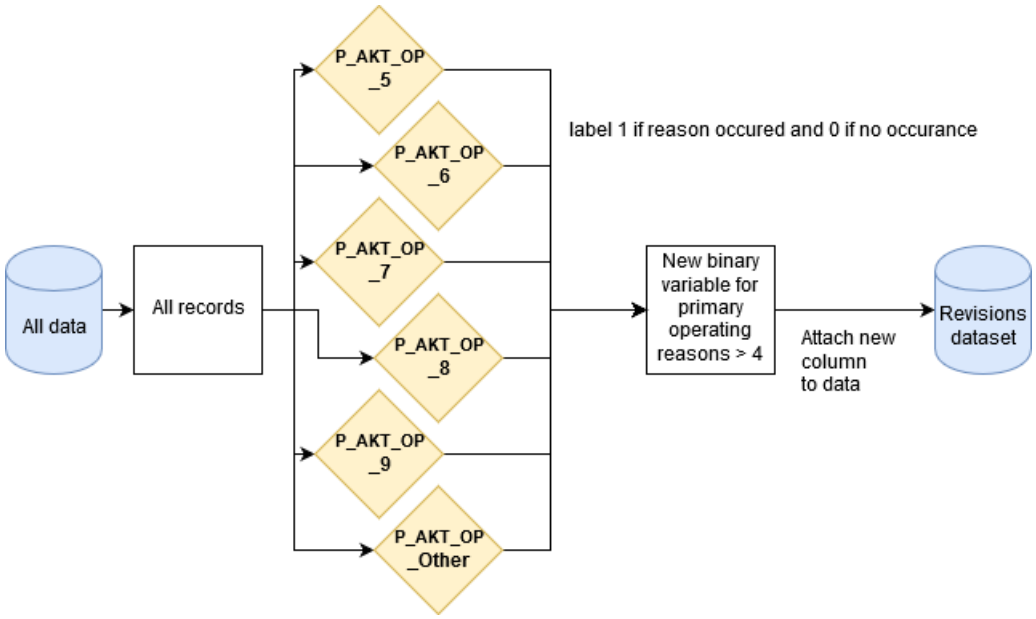

Figure 5.23: Steps taken to add a new column for representing reasons for primary surgery 4-10 in one column as they only appear in a minor selection of records.

Dependent features

For target variables there are the possibility of forecasting a length of survival before revision in exact survival years by *regression*, there is also the current binary revision indicator. However, for binary classification to be meaningful additional context is necessary. Therefore, approaching the issue of classifying an outcome after primary hip replacement surgery is performed through two features named *'Term_Binary'* and '*Term_Multi'*, engineered as discussed above. All the three selected outcomes are listed below.

Selected dependent variables:

Survyrs: Describes exact amount of years survival before revision surgery *(existent)*.
Term_Binary: Describes classes of outcome defined by combining the *Survyrs* and *Revision* variables, establishing the feature *'Term_Binary'* with two possible outcomes dependent on amount of time from last surgery to revision surgery. All records with revision was assigned to one of these three categories *(engineered).*
Term_Multi: Describes classes of revision outcome defined by combining the *Survyrs* and *Revision* variables, establishing the feature *'Term_Multi'* with three possible outcomes dependent on amount of time from last surgery to revision surgery. All records with revisions was assigned to one of these three categories *(engineered).*

Distribution of records according to categorical outcomes:

The establishment of the new outcome features required the revision indicator to be positive, as a known outcome had to have occurred, this reduced the dataset to 5538 patient records. The distribution of records within the engineered features for describing outcome groups are quite evenly split between possible classes. According to the binary outcome variable it is split 54.4% and 45.6%, while the multinomial classes all have around 30-35% of surgery records (Figure 5.24). There is no clear *skewed distribution* of class membership with either the binary or multinomial outcome feature, thus the records are balanced between possible survival outcomes.

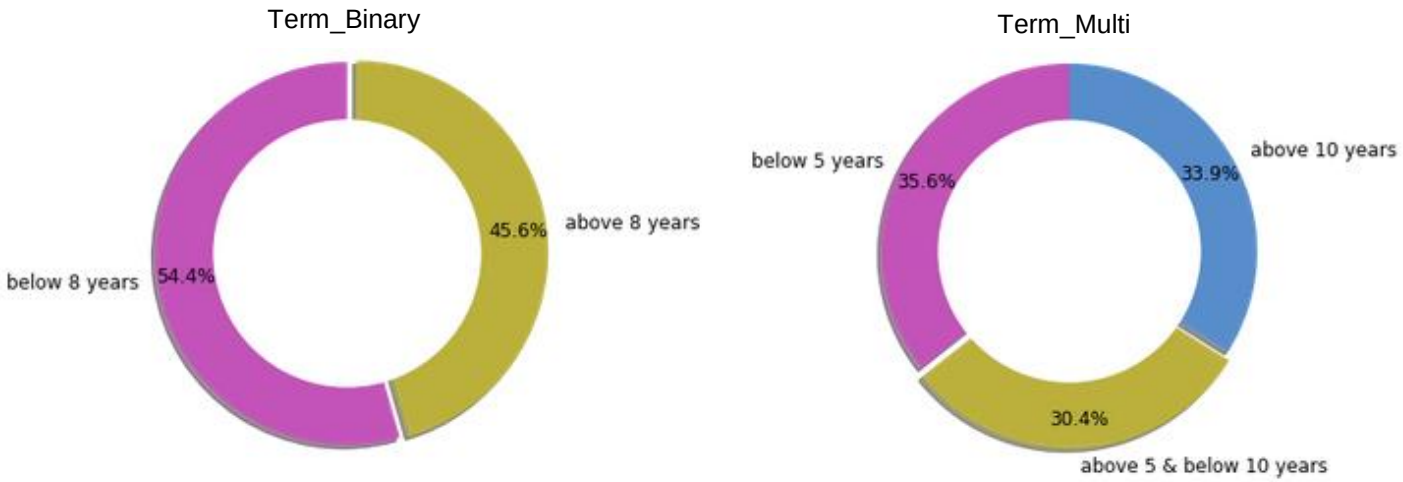

Figure 5.24: Distribution of all revision records according to the newly engineered dependent variables.

6. Modelling

The process of modelling in this thesis is done by applying existent learning algorithms that were suitable for the tasks. The goal is to explore perspectives in predicting an outcome made possible by the establishment of quality registries in *hip arthroplasty*. All perspectives on outcomes are wrapped in the context of *time* in this modelling exercise and is in total three modelling tasks. They are *classifying revision before or after 8 years from primary surgery* and *classifying revision before five years, between* five and *ten,* or *after 10 years from primary surgery.* Furthermore, as a last task predicting *exact survival years* was done to see how much the results would deviate from the exact answer.

The modelling process applies learning algorithms from the *Scikit-learn* library to serve as baseline and for comparing results. Two of the tasks are classification problems, *Logistic regression*, *Random forest classifier*, and *Multi-layer perceptron classifier* were used for modelling these problem spaces as they are applicable to both binomial and multinomial classification (Varoquaux, 2015). Predicting the exact outcome in survival years is a regression task and is performed by *Multiple linear regression*, also from *Scikit learn* (Varoquaux, 2015).

6.1 Approach

The approach details how the process of modelling was performed through three perspectives on predicting a type of outcome overall in this chapter. These perspectives can be illustrated by asking three questions:

1. Will revision surgery be necessary before or after 8 years after primary surgery?
2. Will revision surgery be necessary before 5 years, before 10 years, or later than 10 years after primary surgery?
3. What might the exact length in device survival years until revision surgery be?

The questions were answered through three separate attempts with varying use of features known in a preoperative setting, the features are organized as:

- Patient specific features
- Prosthetic device specific features
- Primary surgery reasons specific features

The features are used in the manner detailed in Figure 6.1, always retaining the patient specific features, and combining them with either device specifications, primary surgery reasons, and with both.

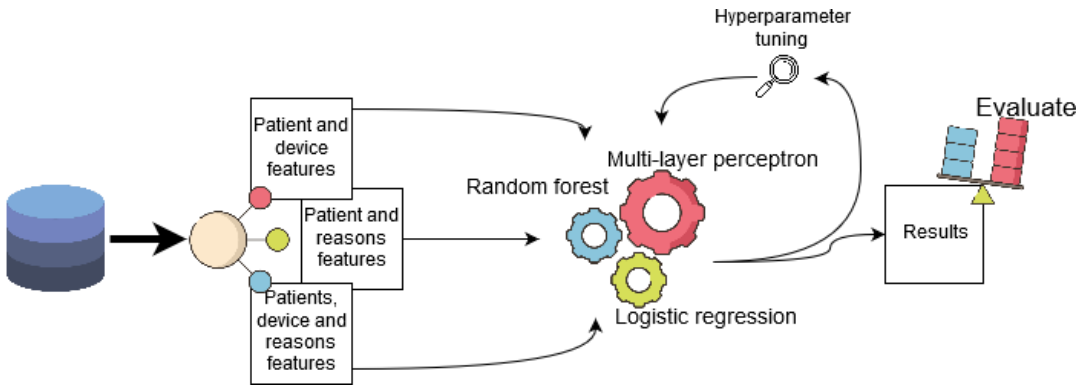

Figure 6.1: Process structure for modelling with different segmentations of preoperative features performed by separating data into different sets and testing different learning algorithms and recording the results.

The process of segmenting data into selections with separate use of features is done to see whether manifestations of consequence or importance would emerge.

6.2 Pre-processing

Feature selection

Feature selection was partly completed through exploring in the previous chapter, three features were selected as *dependent variables*. For a final examination of independent features, *univariate feature selection* was done to see how variables related to their respective outcome. As two of our outcomes concern classes and lastly a continuous outcome, two measures were used to find the importance of features. *F-score* was used to test feature for the regression task and *Chi*^{2}*-score* for the classification tasks (Buitinck, 2013) (Varoquaux, 2015). Results are presented in Figure 6.2.

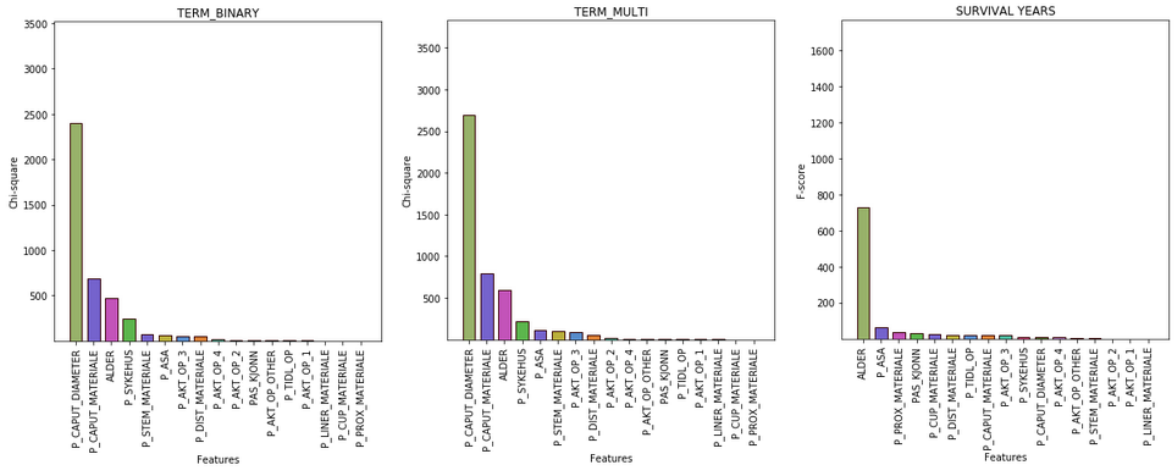

Figure 6.2: Results from feature evaluation with SelectKBest module in Scikit-learn library with Chi² for testing features for classification tasks and f-score for the regression task (Varoquaux, 2015).

The influence on outcome features overall is sparse, with only a few indicating any significant result. There is some deviation in which features are related, the *size* and *material* of the *caput* is the most relevant with binary and multinomial outcome. In case of predicting the exact survival year, the *age* and *patient condition* at surgery (*ASA-class*) is the most significant. All features were included.

Binary encoding

For handling multinomial values in the input feature vector, the *Scikit-learn* module *one-hot encoding* (Varoquaux., 2015) was used to transform the variables on *materials* to a binary matrix representation. This is done to prevent the model from interpreting values on a scale of 0 – 12 as ordinal.

Standardization

The *Scikit-learn* module *StandardScaler* was used to transform the feature vectors to standardize the continuous features, binary representations remain unaltered (Varoquaux., 2015).

Cross-validation

Scikit-learns *cross-validation* module was used for hyperparameter tuning and validating the model on different sections of the data. For the first the number of folds was set to 5, for the latter it was set to 10 (Claesen, 2015).

6.3 Binary revision classification

The binary classification was used to see if revision surgery might be expected before or after a threshold at *eight* years. It is the one out of all three with the broadest defined target, records have an outcome of either class 0 or 1, 0 is before and 1 is after the threshold.

6.3.1 Patient and device features

Cross-validation

Results from cross validation with *patient* and *prosthetic device* specific features is listed in Table 6.1**,** detailing the three classifiers on binary classification over ten folds.

Table 6.1: Cross validation results table with accuracy score from the three models over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	*0.59*	*0.54*	*0.66*
2	*0.55*	*0.52*	*0.57*
3	*0.74*	*0.76*	*0.74*
4	*0.68*	*0.67*	*0.67*
5	*0.63*	*0.62*	*0.66*
6	*0.62*	*0.61*	*0.64*
7	*0.58*	*0.61*	*0.62*
8	*0.59*	*0.60*	*0.60*
9	*0.64*	*0.63*	*0.67*
10	*0.62*	*0.61*	*0.64*

The base line *Logistic regression* model has an overall average accuracy across *ten* folds at 0.624, the *Random forest* model an average of 0.617, while the *MLP* model scores the highest average over all ten folds at 0.647 in accuracy. The results indicate a lack of *variance* in performance across different sections of the data, and the level of *error* appear to be quite high as the accuracy persistently stay below a score of 0.70, only peeking above in one out of ten folds.

Receiver Operator Curve (ROC)

The ROC curve, and area under the curve show a similar result as seen in the overall accuracy, with the *Logistic regression* as the least effective with an area score at 0.718, and *Random forest* and *MLP* with a score at 0.726. The curve is not distinct in any of the three classifiers and show no clear difference across classifiers on patient and device features (Figure 6.3).

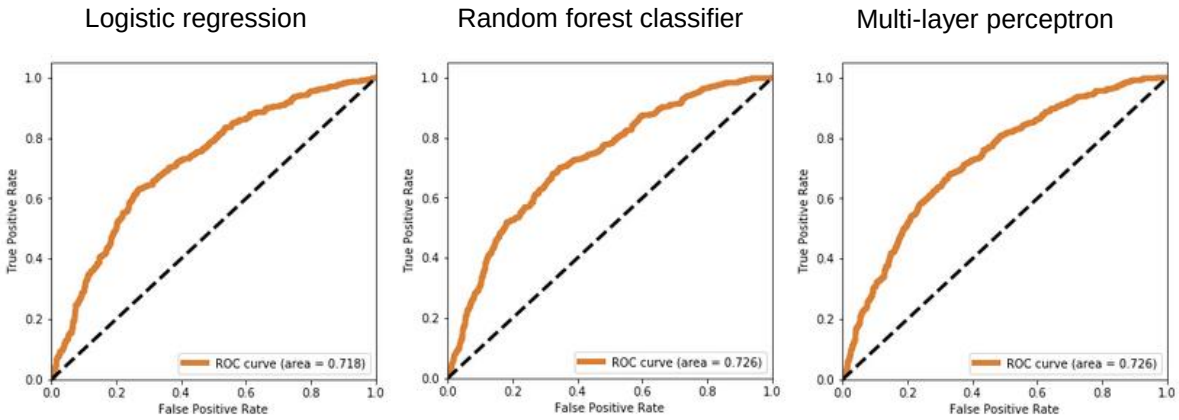

Figure 6.3: Illustration of ROC curve across the three models with patient and device features.

6.3.2 Patient and primary surgery reason features

Cross-validation

Results from cross validation with *patient* and *primary surgery reason* features is listed in Table 6.2, detailing the three classifiers on binary classification over ten folds.

Table 6.2: Cross validation results table with accuracy score from the three models over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	*0.58*	*0.61*	*0.61*
2	*0.57*	*0.58*	*0.56*
3	*0.72*	*0.71*	*0.73*
4	*0.67*	*0.66*	*0.67*
5	*0.64*	*0.68*	*0.67*
6	*0.62*	*0.63*	*0.63*
7	*0.58*	*0.64*	*0.64*
8	*0.59*	*0.61*	*0.62*
9	*0.64*	*0.68*	*0.68*
10	*0.62*	*0.63*	*0.64*

The *Logistic regression* base line has an average accuracy at 0.623 and is surpassed by the *Random forest classifier* at 0.643. The state-of-the-art *MLP* has an average of all ten folds of 0.645 and surpasses the rest by an insignificant margin. The results indicate little or no improvement, with a similar lack of *variance* in performance across ten folds. The error rate appears to be quite high as the accuracy only peeks above 70 in one out ten folds, similar to was observed previously.

Receiver operator curve

The ROC curve, and area under the curve score is similar across classifiers, with the *Logistic regression* as the least effective with an area score at 0.718, *MLP at 0.723* and *Random forest classifier* with an area score at 0.725. The curve has no clear distinctions on visual inspection, although a larger similarity in curvature between *Logistic regression* and *MLP* across all records (Figure 6.4).

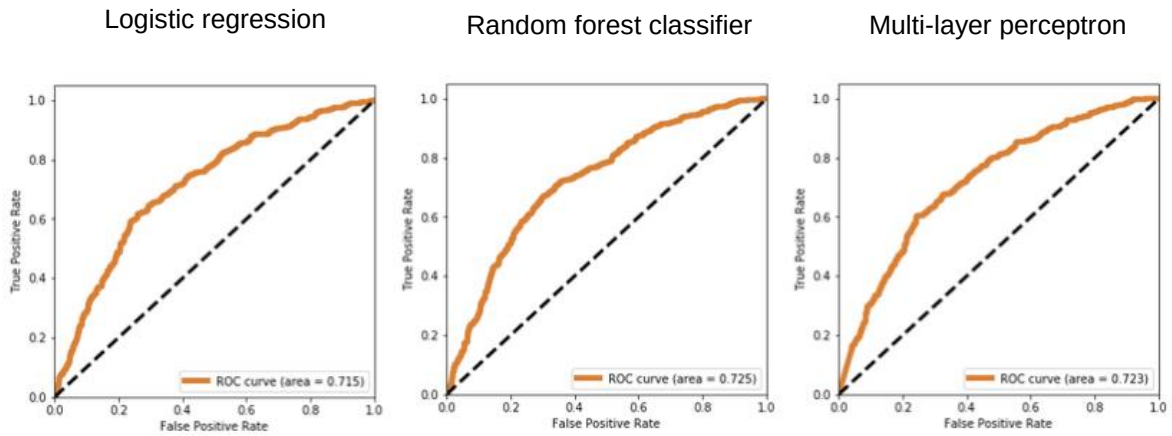

Figure 6.4: Illustration of ROC curve across the three models with patient and primary surgery reason features.

6.3.3 Patient, device and primary surgery reason features

Cross-validation

Results from cross validation with *patient*, *prosthetic device*, and *primary surgery reason* features is listed in Table 6.3**,** detailing the three classifiers on binary classification over ten folds with all three feature categories.

Table 6.3: Cross validation results table with accuracy score from the three models over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	*0.58*	*0.56*	*0.59*
2	*0.55*	*0.52*	*0.64*
3	*0.75*	*0.75*	*0.73*
4	*0.68*	*0.68*	*0.67*
5	*0.64*	*0.62*	*0.66*
6	*0.63*	*0.62*	*0.65*
7	*0.59*	*0.60*	*0.63*
8	*0.59*	*0.59*	*0.61*
9	*0.65*	*0.62*	*0.67*
10	*0.62*	*0.60*	*0.68*

The *Logistic regression* base line has an average accuracy score at 0.628, surpassing the *Random forest* classifier with a score of 0.616. While the *MLP* classifier outperforms the other with an accuracy at 0.653. The results indicate no consequential improvement, with a similar lack of *variance* in performance across ten folds with increased feature space. The *error rate* appears high as the accuracy largely stay persistent within the *.60-70-range* across ten folds, only peeking above or below a $< .5$ accuracy.

Receiver operator curve (ROC)

The ROC curve, and area score follow a similar suit as prior, with the *Logistic regression* as the least effective with an *area under the curve* score at 0.715, *Random forest classifier* with a score of 0.722, and *MLP* surpassing the latter by an insignificant 0.002. The curve has no abrupt changes across the spectrum, although the *Random forest classifier* has a slightly more varied performance (Figure 6.5).

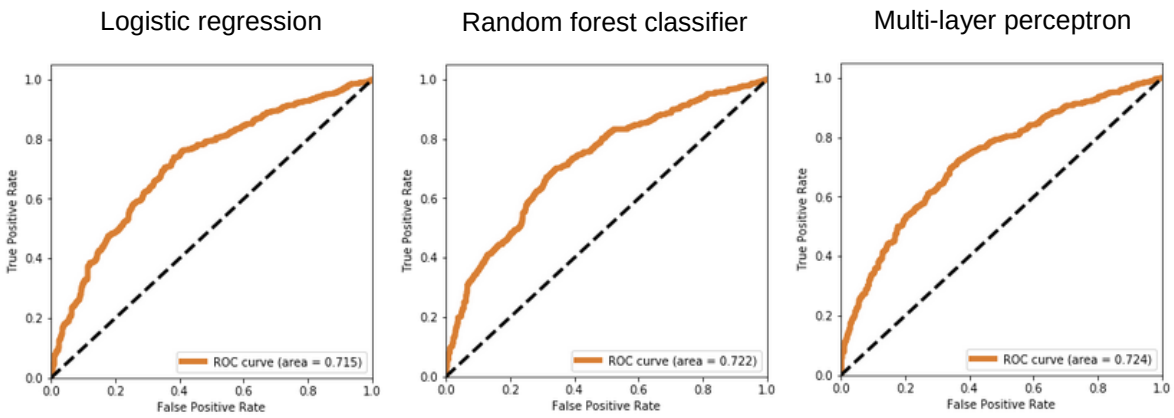

Figure 6.5: Illustration of ROC curve across the three algorithms with patient, device and primary surgery reason features.

6.4 Multinomial revision classification

The multinomial classification attempts to predict if revision can be expected before *five years*, between *five and ten*, or after *fifteen* years. This task has a narrower target than what was attempted in *Section 6.3*, increasing possible outcomes to three. The records can have an outcome of either class 0, 1, or 2.

6.4.1 Patient and device features

Cross-validation

Results from cross validation with *patient* and *prosthetic device* features are listed in Table 6.4, detailing the three classifiers on multinomial classification over ten folds.

Table 6.4: Cross validation results table with accuracy score from the three models over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	*0.41*	*0.40*	*0.43*
2	*0.51*	*0.51*	*0.52*
3	*0.55*	*0.56*	*0.53*
4	*0.47*	*0.48*	*0.47*
5	*0.49*	*0.48*	*0.50*
6	*0.46*	*0.47*	*0.46*
7	*0.42*	*0.44*	*0.44*
8	*0.43*	*0.45*	*0.45*
9	*0.47*	*0.47*	*0.47*
10	*0.47*	*0.47*	*0.48*

The *Logistic regression* classifier has an average accuracy at 0.468, a considerable drop from the performance seen overall in binary classification. *Random forest classifier* scores 0.43, while *MLP* outperforms the rest with 0.475. Overall, the models have no larger fluctuation in performance, and stay somewhat stable within 0.4-0.5, only peeking above on similar folds as the others. The results indicate a lack of *variance* in performance with only small changes in generalizing model performance across different sections of the data. The level of *error* appears to be high as the accuracy persistently stays below .50.

Receiver operator curve

The ROC curve, and area score is measured as a one-vs-all, indicating a slightly more optimistic representation. Results are similar across the board, the *Logistic regression* getting class 0 and 2 wrong, and more correct classifications of class 1. Overall, the micro-/macro-average show *MLP* performing better than the rest by a minor 0.01. The classifications are skewed toward class 0 and 2 (Figure 6.6).

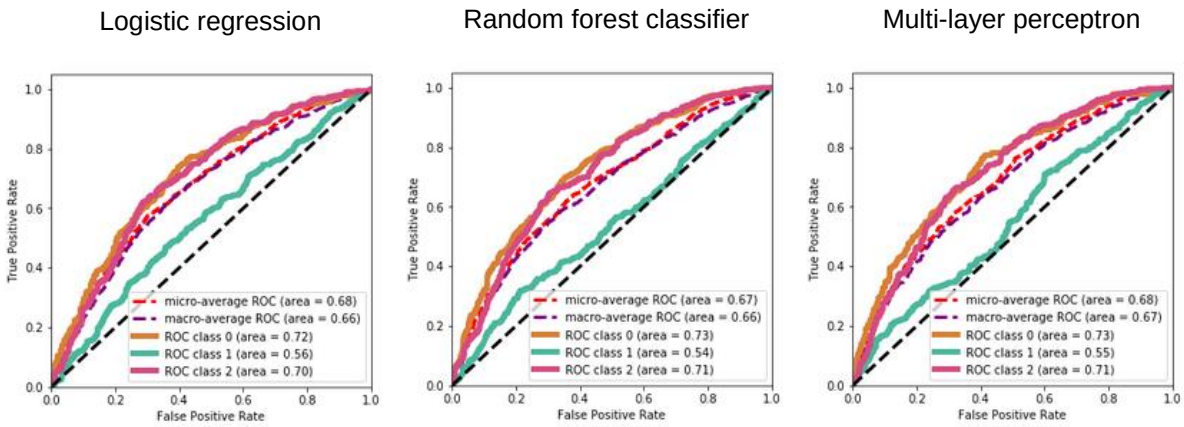

Figure 6.6: Illustration of ROC curve across the three models with patient and device features.

6.4.2 Patient and primary surgery reason features

Cross-validation

Cross validation with *patient* and *primary surgery reason* features is listed in Table 6.5, across the three classifiers on multinomial classification.

Table 6.5: Cross validation results table with accuracy score from the three models over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	0.43	0.42	0.44
2	0.52	0.54	0.53
3	0.55	0.55	0.54
4	0.47	0.47	0.45
5	0.49	0.48	0.50
6	0.46	0.48	0.47
7	0.43	0.45	0.44
8	0.44	0.46	0.46
9	0.47	0.47	0.47
10	0.46	0.48	0.48

Starting from left, the *Logistic regression* still scores less overall, averaging all results to 0,472, followed by *MLP* also on 0.478, and *Random forest* having an accuracy at 4.8, performing slightly better than the others. Although, overall the scores are low, the results indicate low variance with only minor alteration in performance throughout ten folds. Similar to previous results, the *error* appears to be high and stable as the accuracy is persistently low across all folds.

Receiver Operator Curve (ROC)

The ROC curve with surgery reason details does not cause any apparent consequential changes on performance. The *Logistic regression* scores best on classifying class 1, while *Random forest* classifier on inspecting the curve does worse than the rest in class 1. Overall, the micro-/macro-average scores indicate no clear difference in performance between the classifiers as the results are identical (Figure 6.7).

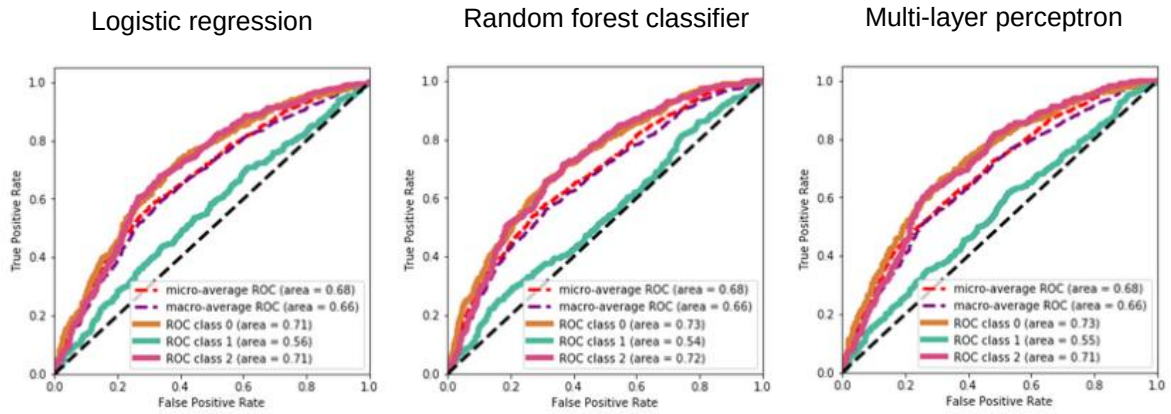

Figure 6.7: Illustration of ROC curve across the three models with patient and primary surgery reason features.

6.4.3 Patient, device and primary surgery reason features

Cross-validation

Results from cross validation with *patient*, *prosthetic device* and *primary surgery reason* features are listed in Table 6.6, with the three classifiers on multinomial classification. All selected features were used here.

Table 6.6: Cross validation results table with accuracy score from the three algorithms over ten folds.

FOLD	LOGISTIC REGRESSION	RANDOM FOREST	MLP
1	*0.42*	*0.38*	*0.43*
2	*0.52*	*0.51*	*0.52*
3	*0.55*	*0.58*	*0.54*
4	*0.46*	*0.46*	*0.45*
5	*0.49*	*0.47*	*0.49*
6	*0.46*	*0.47*	*0.46*
7	*0.44*	*0.44*	*0.44*
8	*0.44*	*0.45*	*0.46*
9	*0.47*	*0.47*	*0.47*
10	*0.46*	*0.47*	*0.48*

Increasing the number of features to include all selected categories and the scores are still similar to previous sections, with no observed consequential changes in performance. The overall average accuracy is between 0.47-0.48 on all classifiers, with no signs of larger *variance* in results, though *Random forest* varies most in accuracy over ten folds. The results indicate low variance with only minor change in accuracy throughout ten folds. Furthermore, on increasing number of features the *error rate* is still high and stable as the accuracy is persistently low, although, there is a slight (rather insignificant) improvement in overall accuracy.

Receiver Operator Curve (ROC)

Results from the ROC curve show some decrease in performance in the *Random forest* classifier, specifically in classifying class 1. The results from *Logistic regression* and *MLP* do not show any noticeable change in performance. The micro-/macro-average scores indicate no changes of importance, *MLP* exceeds the rest by a small margin, a phenomenon observed previously (Figure 6.8).

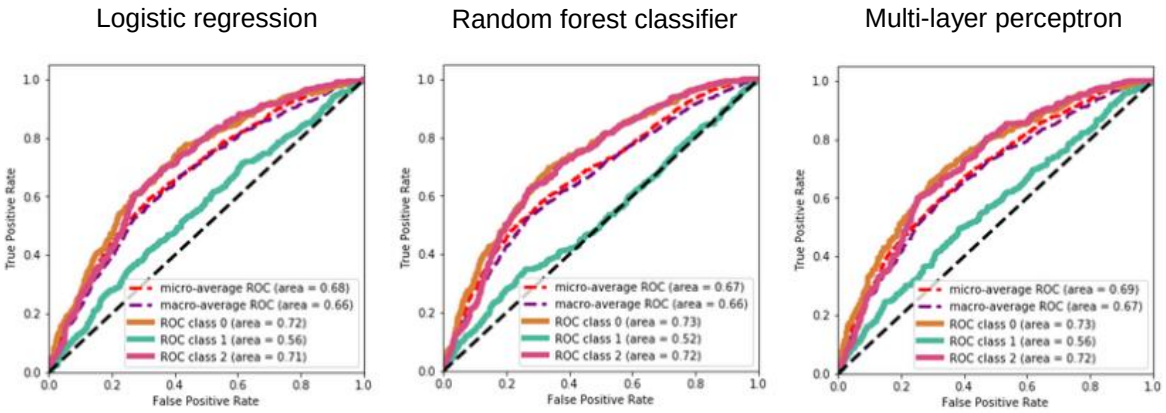

Figure 6.8: Illustration of ROC curve across the three algorithms with patient, device and primary surgery reason features.

6.5 Predicting exact device survival year

The regression task of predicting the length of survival of a prosthetic device until revision surgery is the narrowest of the three target outcomes, a much more specific result than what was done in classification. Although a target with a much smaller chance of a correct outcome, it provides valuable insight into how the predicted value deviated from the truth, and a clearer picture on how it responds to larger variations in device survival length.

6.5.1 Patient and device features

Cross-validation

Cross validation with *patient* and *prosthetic device* features is listed in Table 6.7, listing four metrics describing the amount of error and variance, and overall performance.

Table 6.7: Cross validation results table with regression metrics from the three algorithms over ten folds.

FOLD	R^2	MAE	MSE	RMSE
1	*-1.417*	*646*	*2.7*	*3.7*
2	*0.008*	*3.4*	*17.3*	*2.9*
3	*-8.453*	*7001*	*2.6*	*4.3*
4	*0.134*	*4.6*	*30.7*	*4*
5	*0.151*	*4.4*	*29.3*	*4*
6	*0.14*	*4.3*	*28.7*	*3.9*
7	*0.121*	*4.7*	*34.4*	*4.2*
8	*0.119*	*4.6*	*31.1*	*4*
9	*0.145*	*4.7*	*33*	*4.2*
10	*0.151*	*4.6*	*33.2*	*3.8*

The results show a larger occurring rate of error between the predicted and known *ground truth*, there is some stability in the magnitude of total errors across ten folds, though there is also range of larger fluctuations using the *device specific data* as the only additional features. Predicted outcomes have an average error below five years deviation from the truth, although the error in size is not the largest in most predictions but the number of errors is many in quantity. Overall performance by the R^2-score indicates an unreliable model with almost no chance of predicting the exact survival length.

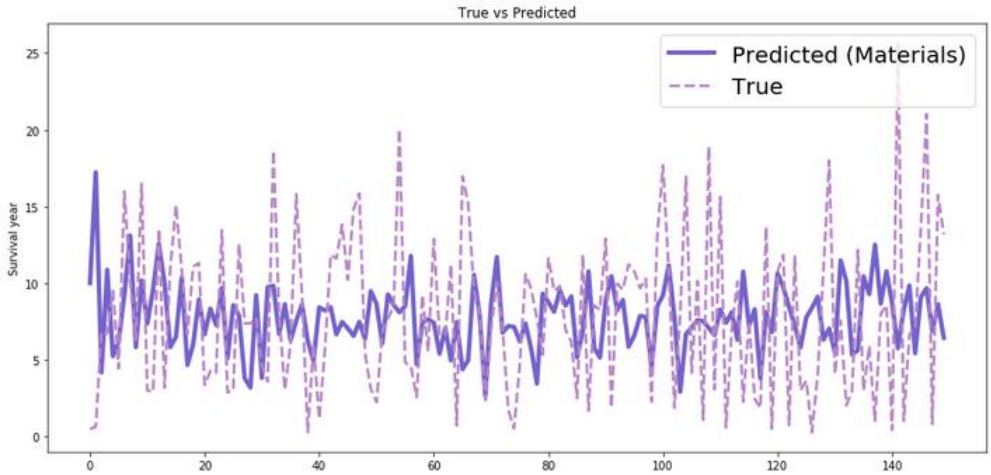

Figure 6.9: Illustration of a selection of predictions vs the known true outcome in survival years with patient and device materials features.

Visual inspection of results from Figure 6.9, showed the model largely remain inside a range between *three* and *thirteen* years, seldom predicting above *thirteen* years and below *three*. The model performs poorly on records with a significant deviation from the average survival length when using the *patient and device features*.

6.5.2 Patient and surgery reason features

Cross-validation

Cross validation with *patient* and *primary surgery* features is listed in Table 6.8, describing the amount of *error* and *variance*, and overall performance.

Table 6.8: Cross validation results table with regression metrics from the three models over ten folds.

FOLD	R^2	MAE	MSE	RMSE
1	*0.13*	*3.3*	*16.9*	*3.2*
2	*0.038*	*3.3*	*16.8*	*2.9*
3	*0.112*	*4.2*	*27.7*	*3.7*
4	*0.134*	*4.5*	*30.7*	*4*
5	*0.157*	*4.4*	*29.1*	*3.9*
6	*0.14*	*4.3*	*28.8*	*3.8*
7	*0.128*	*4.7*	*34.1*	*4.2*
8	*0.117*	*4.6*	*31.6*	*4.1*
9	*0.148*	*4.8*	*32.9*	*4.3*
10	*0.146*	*4.6*	*33.4*	*3.9*

The mean square error indicates a larger rate of errors occurring, and a low rate of *variance* in number of errors across separate test sets. Overall specificity is quite poor, although the score is improved slightly by changing from *device materials* to *primary surgery reasons* as the additional features. The fluctuations in results are minimal, within a difference at +/-1 in total summed error and root mean square error across the spectrum of testing data.

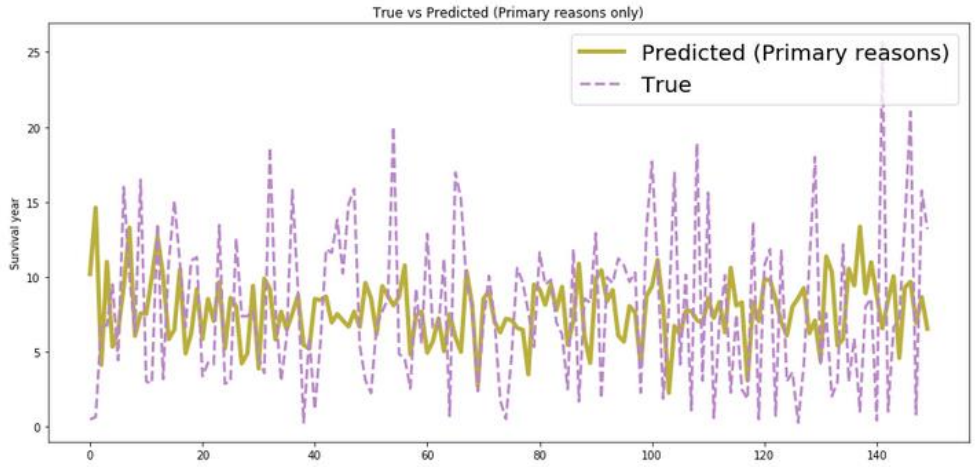

Figure 6.10: Illustration of a selection of predictions vs the known true outcome in survival years with patient and primary surgery reason features.

Visualizing predictive results are in Figure 6.10. A similar phenomenon repeats itself as the mean square error rate is below five years in average size, and the larger errors appear at a decent gap in distribution while most frequent errors are smaller in size. Generally, the model is poor in performance and still has large amount of errors and produces results that are not varying much in predicted values.

6.5.3 Patient, device and surgery reasons features

Cross-validation

Cross validation with *patient*, *device*, and *primary surgery* features are listed in Table 6.9, describing the amount of *error* and *variance*, and overall performance.

Table 6.9: Cross validation results table with regression metrics from the three models over ten folds.

FOLD	R^2	MAE	MSE	RMSE
1	0.012	3.6	19.2	3.4
2	0.02	3.4	17.2	3.1
3	0.138	4.1	27.1	3.7
4	0.136	4.6	30.1	4
5	0.161	4.4	29.1	3.9
6	0.131	4.7	29	3.9
7	0.128	4.6	34	4.2
8	0.125	4.6	31.3	4.1
9	0.168	4.9	32.8	4.3
10	0.149	4.6	33.4	3.8

Results reveal no significant changes in model performance by increasing the feature space to include all categories of selected features. This increases the number of features significantly but does not change performance of the model in any of the aspects of overall error rate, root mean size of errors, and frequency of larger deviations. The change is barely noticeable, although there are more errors and less variance by increasing number of features, there are no distinct changes appearing to be of consequence or importance.

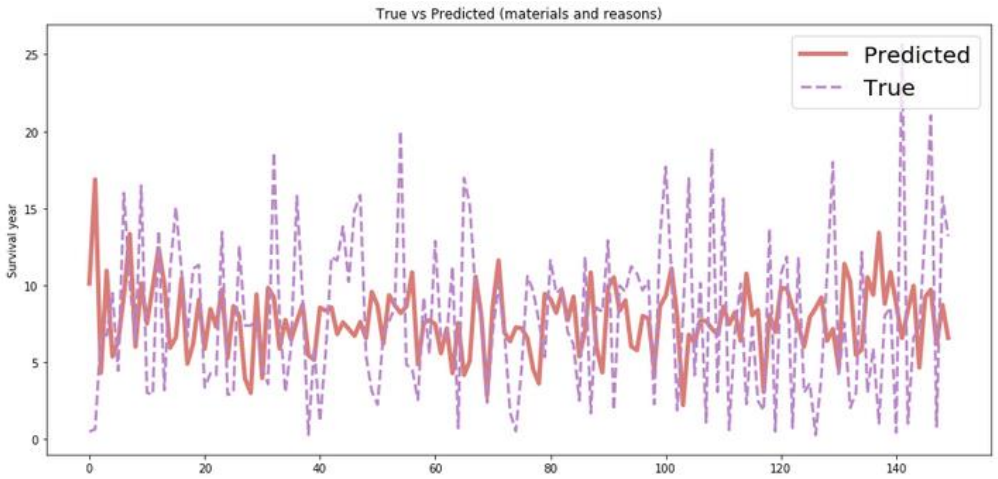

Figure 6.11: Illustration of a selection of predictions vs the known true outcome in survival years with patient, device materials and primary surgery reason features.

Visualizing gave a similar illustration of performance seen in the two previous sections (Figure 6.11), there are no clear changes. The model struggles to grasp errors with a longer than average survival length (above 8-10 years deviation from the mean) with no real chance of getting within a short range of the *ground truth*. On smaller errors the model can get somewhat close to the truth, however the mean errors in predictions are still approximately *four* years.

7. Results

The chapter is a broader reflection on the results achieves in the last chapter and summarizes the exploration and its aftereffects. Further, a review of the results from modelling the bi- and multinomial classification tasks, and a brief look at the prediction of exact survival year of a prosthetic device.

7.1 Exploration

In this thesis data exploration was done to gain a sense of situational awareness by uncovering more about what it could describe about the surgery records, and the differences and similarities between the groups of arthroplasty patients. The goal of the *exploratory data analysis* was to locate perspectives on predictive modelling, therein potential *dependent* and *independent* variables as outcome and explanatory features. Several aspects of the data were examined. Regarding missing values most columns were present, although among important variables such as *patient health status* and *caput diameter* there was a majority missing values. In depicting missing values, the relationship between presence of positive values and neutral 0, registered absence of a value, was briefly examined prior to clustering the data. In reasons for *requiring primary surgery* the majority was one type, i.e. *coxarthrosis*, although the reasons are described dichotomously and not restricted to only one at the time. The presence of other reasons was very low. The variables on materials used have a wider variety in certain device parts, there is however a minority in variety as one of the product types was represented at a much larger scale than the others (*gold records*).
Clustering the data gave insight into the larger similarities and the minor differences in the case of this dataset, in *age* the records were largely similar across different clusters, as well as by *product usage*. Considering *gender* and reoccurring survival outcomes from clusters there was no clear indication of any group exposed to certain risks. The records are approximately two thirds female majority. Visualising clusters by revision indicated there is no disproportionate distribution of records according to gender, as this characteristic is a repeating phenomenon in produced clusters. There is minimal variety in the dataset, even though there are several materials represented, the majority of prosthetic devices fall within a few combinations of these materials

Outcomes

The variables suitable for outcomes were the *length of survival of the device* and an *indicator for revision surgery*. Most records had no occurrence of revision surgery, meaning the device was still functioning. The survival year was then difficult to interpret as a definite answer to

how long a device survived until revision, as no revision surgery has been required. There are interesting outcomes among variables in the dataset, *reasons for requiring revision surgery* and the separate adverse events could have been approached as a categorical classification if combined with the revision indicator. However, in this selection of data the variation in adverse events is sparse, most are largely aseptic loosening of the stem and/or cup, and no signs were seen among groupings of records indicating a larger presence of any one *type of reason for revision*.

Target features

In *Section 5.4*. two dependent features for classifying an individual patient outcome into a class based on how long the expected survival length of the prosthetic device would be were engineered. The first feature detailed a class distribution based on a threshold at *eight years* by labelling records with a known revision status as a binary classification problem. The basis for the classification is the recorded length in years between primary and revision surgery. The second detailed a three-way split, distributing records into a class by two thresholds at *five years* and *ten years*. Surgery cases of patients with a known length below *five years* became one class, between *five* and *ten year* became the second class, and above *ten years* became the third class. By labelling records with a known revision within any of these classes the number of records was reduced significantly. This excluded all records with no positive indication of revision surgery, thereby removing patients deceased before anything went wrong and those where the outcome in years is still unknown due the patient moving away. As a last prediction the continuous feature detailing *exact length of survival in years* was included by reducing the dataset to only records where a revisions surgery was known to have occurred after the given length in device survival. This served as an interesting look towards how much a prediction would deviate in number of years as the dependent feature differed from the two classification tasks and is given years, thus making it is interpretable to most.

7.2 Modelling

Distribution of records among the new features indicated the population was not largely skewed toward one or the other class (Section 5.6.1):

- According to the binary revision variable before or after *eight years*, the population was split to *54.4%* and *45,6%*, suggesting the majority of those who have revision surgery among this data have it before eight years passed.
- According to the three-way split multinomial variable the population was distributed with *35,6%* with revision before *five years*, *30,4%* between *five* and ten years, and *33,9%* after *ten years*.

Binary classification

The binary classification considered whether a patient would require revision surgery and replacement of a prosthetic device before or after *eight years* from primary surgery. The features used were based on information available prior the actual surgery, and contained variables on the patient, unique materials in the device, and the reasons for requiring the implant, i.e. types of arthritis and fractures. The distribution of revision records by the known outcomes showed nearly equal distribution between before or after eight years. This is comparable in distribution to the predicted result from Chapter 6 which had a similar distribution in the *Multi-layer perceptron classifier* with *53.5%* below and *46.5%* above eight years (Figure 7.1).

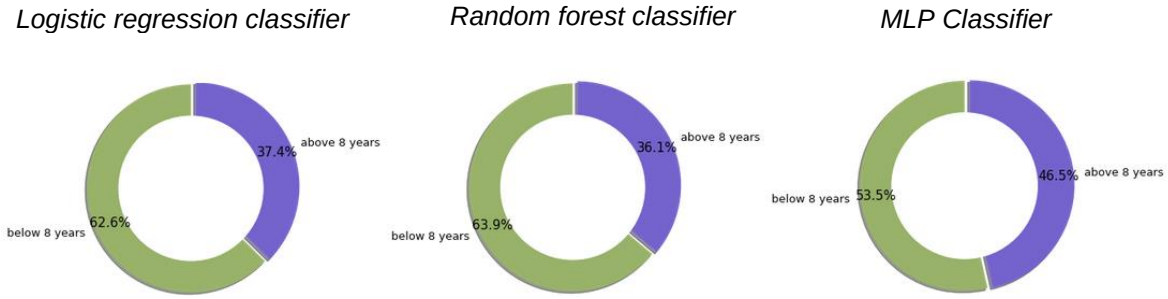

Figure 7.1: Visualization of class distribution by each model with binary target and all features (patient, prosthetic device and surgery reasons features).

Although the distribution is similar, the accuracy score was only between *60-70%* and a minority of predictions were either *false positive* or *false negative* (Figure 7.2). The *MPL classifier* was slightly better to other models, however it was only by a small margin. The results from the confusion matrix showed *MLP* was slight better at correctly predicting occurrence of revisions before eight years, and slightly worse on predicting revision after eight years. The *Random forest classifier* had a correct classification rate of class 0 and 1 at *43.7%* and *23.3%*, respectively. Moreover, *12%* and *20.2%* in incorrect classification for class 0 and 1, respectively. The best model was MLP, with a *sensitivity* at 0.72 and a *specificity* at 0.66, however the random forest classifier was not far behind with sensitivity and specificity at 0.77 and 0.53, respectively.

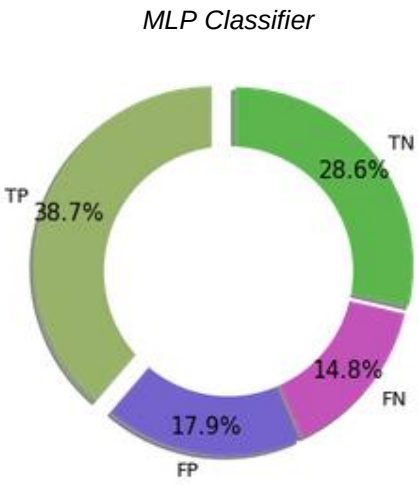

Figure 7.2: Visualization of confusion matrix from MLP with all data.

Multinomial classification

The multinomial classification tried to classify whether a patient would require revision surgery and replacement of a prosthetic device before *five years*, between *five* and *ten years*, or after *ten years* from primary surgery. The distribution of revision cases by the known true outcome showed a close to equal distribution between the three classes, however, the predicted results from Chapter 6 had a different distribution. The *Random forest classifier* and *Logistic regression classifier* had a similar distribution of classifications between them, with only a small minority of records expected to require revision surgery after *five years* and before *ten years* from primary surgery (Figure 7.3). The *MLP classifier* have more class 1 (between five and ten years) outcomes as a result, but the accuracy was not found to be consequentially better.

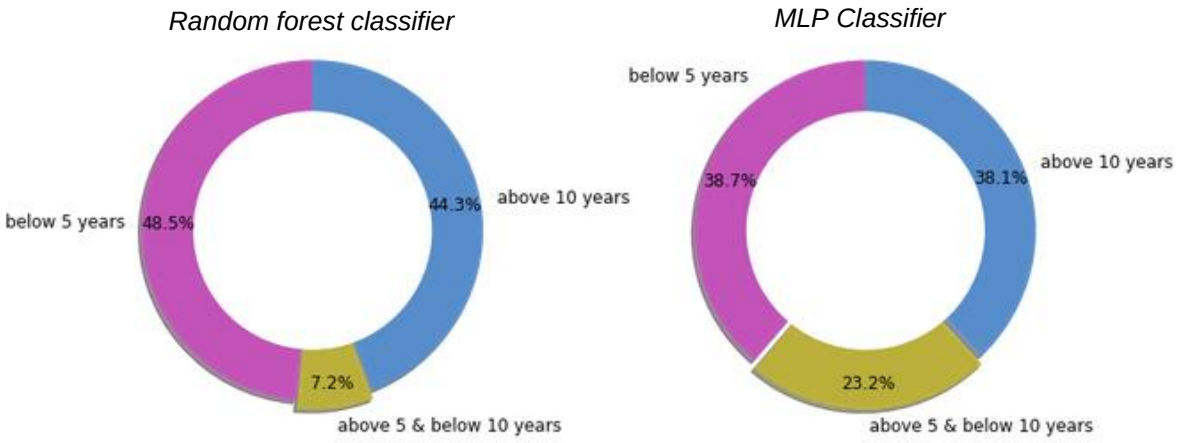

Figure 7.3: Visualization of class distribution with three classes by two models, random forest and MLP, and all selected features (patient, prosthetic device and surgery reasons features).

The *sensitivity* and *specificity* of the multinomial classification models were calculated as *one-vs-the rest* for each possible outcome class. The *MLP classifier* had a *sensitivity* and *specificity* at 0.71 and 0.55 on classifying class 0. Class 1 had a *sensitivity* at 0.77 and *specificity* 0.25, while class 3 resulted in 0.69 and 0.53, respectively. The *sensitivity* measures how well the models perform at detecting an event of a certain class (Geron, 2017, p.91), meaning the models are able to correctly classify approximately 70% of records of class 1 correctly. However, the *specificity* measures how precise the model classifies a certain outcome class (Geron, 2017, p. 91). Meaning considering the results above the model is not very precise in correctly classifying any of the three classes, with most correct classifications being of class 0 (below 5 years device survival).

A phenomenon observed across all three classes was a *specificity* lower than the *sensitivity*, with class 1 at 0.25 *specificity* indicating a larger selection of incorrect classifications of between *five- and ten-years* device survival before revision.

The *Random forest classifier* had a *sensitivity* and *specificity* for class 0 at 0.63 and 0.67, class 1 at 0.93 and 0.09, and class 2 at 0.66 and 0.66. In contrast to the result from *MLP*, the scores are more evenly distributed between the model being good at predicting a certain class and precision for said classifications. In classifying class 1 the model only assigns 7.2% of records as expected to have a device survival between five and ten years before revision. On these records the s*pecificity* score was very low with only 9%, showing a less balanced result than the *MLP classifier*. The *Logistic regression classifier* and *Random forest classifier* has largely similar results.

Predicting exact survival year

The task of predicting the exact survival years of a prosthetic device after primary surgery was done by *multiple linear regression* and through a similar process as with the classification tasks. It was done as three separate exercises, always retaining the patient specific features, and combining them with device materials features, primary surgery

reason features, or both selections of features. The results showed that an exact survival year was a difficult target to predict, and the model could not get close to the know *ground truth* in most cases. Examining the results with the algorithm trained on separate selections of features display a similar performance with little differences observed in predicted outcomes regardless of the explanatory features (Figure 7.4).

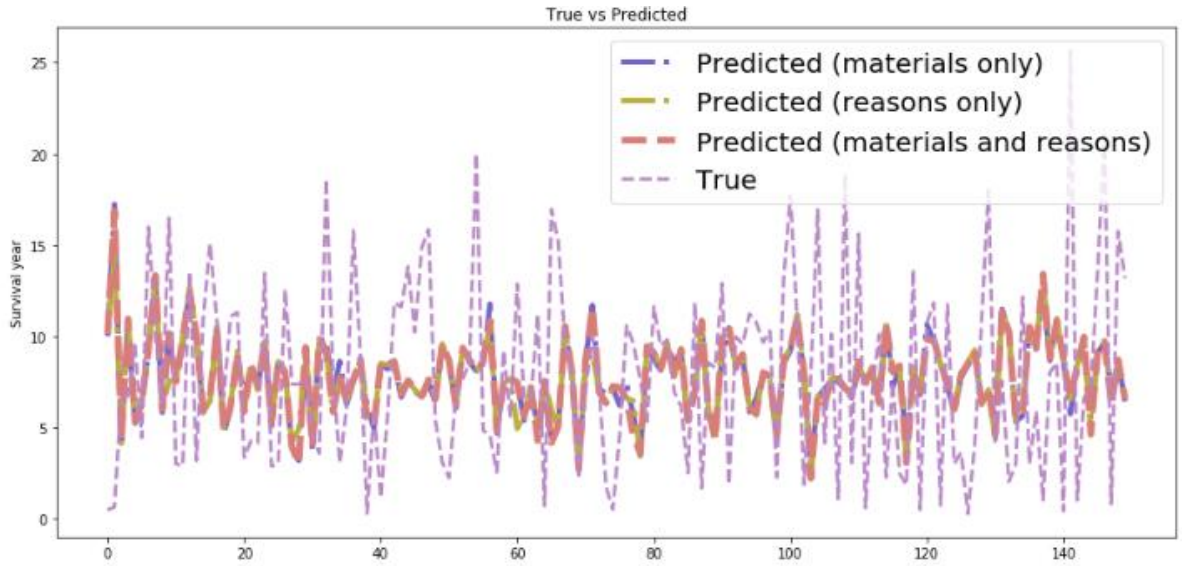

Figure 7.4: Illustration of a selection of predictions from multiple linear regression, employing different sets of features.

The use of different sets of features did not influence the amount of error and variance observed. The *Mean Absolute Error* was at approximately four years across the spectrum, with only minor deviation seen between model performances when changing features.

Similar effect was seen in the *Root Mean Square Error*, with most results from tenfold cross validation giving an error at approximately four years. Across the ten folds the model gave results with little variation in performance by training and testing on different sections of the dataset, meaning their performance experienced little fluctuation in terms of amount of error.

8. Discussion

The research is discussed in this section, starting with the employed methods and what they yielded. This is followed by as brief look at variables ability to predict an outcome, and lastly, the research questions are answered.

8.1 Methodology and methods

This section discusses the Design Science and how it assisted in guiding the research, the essential nature of data exploration in this thesis, and the machine learning methods employed to perform predictive modelling.

Design Science Research

In this project *Design Science* was adopted as a fundamental methodology in guiding and research and development in this thesis. The Design Science research methodology gave a framework on principals of relevance and rigor in design of an artefact in a research context (Section 3.1). In general, the artefact is the establishment and testing of machine learning models for individual outcome prediction in hip arthroplasty. It consists of two parts; the first part of the artefact are the dependent variables selected and engineered from the dataset to make an attempt at prediction feasible (Chapter 5). The second part is the training and testing of predictive solutions by implementing learning algorithm from *Scikit-learn* (Chapter 6). The research questions were introduced in *Section 1.2* and provided crucial direction on how to move forward in exploring the dataset and modelling possible outcome solutions. Having established the goal as prediction of individual patient outcome and a search for the prerequisites to do so helped navigate in decision-making throughout the process. First two questions are concerned with establishing the possibility of implementing the predictive models based on identifying available features and their potential as an outcome. The third, based on the results from the first research questions, then attempts to use the findings to implement prediction models for hip arthroplasty. This conduct of research was helped by follow Design Science providing a structure stopping the work from going either above or below the intended range, in purpose of the research and development. Further, it did as well help in remaining focused on the research aspects in the establishment of an artefact, promoting inclusion of relevant research and the purpose of facilitating knowledge.

Exploration

The *exploratory data analysis* was performed to address the first part of the artefact by finding and establishing the outcome features necessary to perform a modelling task. At the beginning of the thesis not much was known about the dataset except the distribution according to different product types (Section 4.2). Possibilities for testing individual outcome predictions within this dataset was unknown information at the time and had to be discovered. This is where the exploration played a crucial role as it firstly brought forward contextual knowledge about the data itself, i.e. the number of records and how these records are represented by different variables explaining the similarities and differences between them (Chapter 5). In light of the research questions it was necessary to figure out if some dependent variable existed which could be an interesting option for predictions. As well as, if the dataset had the variables necessary to make new dependent features by combining existing ones. Furthermore, it was of interest to locate independent variables which could be used to explain the occurrence of a specific outcome.

The exploration phase was rather invaluable in assessing these variables within the data as dependent and independent features. Specifically, for this study there was a necessity for engineering new dependent features as outcomes to model towards, as the data was sparse in variables describing outcome scenarios for individual patients. The occurrence of an adverse event, where revision surgery and implanting a new device would be required, was found to be the one type of outcome. It also showed this type of outcome could be approached from different perspectives, one is the classification of which adverse event will occur and the other is to approach the issue in a time related sense. The first would predict what type of reason for revision might be, while the latter focuses on the time a revision surgery might become necessary. The first task was found to be more difficult modelling exercise as the data was heavily skewed toward aseptic loosening, similarly it had a majority one type of prosthetic device leaving little variation in the dataset on reasons for requiring revision surgery. The latter, focusing on the time-based nature of the occurrence of revisions surgery, was found to be a much more plausible route as the dataset had a wider variety in when revision occurred in contrast to how it occurred. Therefore, in predicting an outcome a result contextualized by length of survival became the reason for selecting and engineering dependent outcome features. This led to three dependent features being chosen to proceed with in modelling, the exact survival years were selected as one outcome by including records where the listed years detailed an actual length until revision. While two more dependent features were selected by engineering new categorical outcome features. The availability of the exact survival years and an indicator for occurrence of revision surgery made it possible to engineer outcome classes based on either below or above a certain device survival length in years. The revision indicator was of major importance as it made the separation of records possible due to a known true outcome. Therefore, the exploration gave an opportunity to establish a collection of records suitable for learning an algorithm to predict and to evaluate the results by comparing predicted and true values.

Predictive models

This thesis utilized a selection of *machine learning models* for supervised learning (Section 3.2.4), as a means for predicting both discrete dependent variables by assigning records to a class and for predicting toward a continuous dependent variable. The classification predictions towards a discrete target were made possible by the engineering of two variables as a result from the exploration, one binary classification and one multinomial variable (Section 5.6.1).

Binary classification: The results from binary classification showed an accuracy of prediction around 60-70% and were the one of the three predictive modelling tasks with the highest score (Section 6.3). The prediction was done with three different uses of explanatory features; once with only *patient and device specific features*, once with *patient and primary surgery reason specific features*, and one with *patient and both device and primary surgery reason features*. There were minimal differences observed by using different features, the overall accuracy of the classifications did not improve significantly by changing features (Chapter 6). The results from using patient and primary surgery reasons were slightly above the results from using patient and device features, while increasing the feature space to all selected features did not show consequential change in performance. Different features were tested on all three learning algorithms adopted in this thesis, a reoccurring phenomenon was the observation of a slight improvement in the *multi-layer perceptron* classifier compared to the *logistic regression* and *random forest* classifiers.

Multinomial classification: The results from the multinomial classification indicated that the model trained on the dataset used in this thesis did not provide the necessary information for mapping a surgery record to the correct outcome class among multiple options. The accuracy score achieved by the models through testing generalization with cross validation showed the models did not score any better than what a random guess would, only reaching above 50% accuracy on a selection of folds (Section 6.4). The score below 50% suggests the features available in the dataset may not provide the necessary explanatory power to achieve a classification of record according to the multinomial classes. Classification was done with separate use of features, but no consequential change was seen by altering the features the models were trained on (Section 6.4). Similar to the results from the binary classification, the *Multi-layer perceptron classifiers* performed better than the other models provided by the *Scikit-learn* library, although the changes in performance were barely noticeable between all classifiers.

Predicting exact device survival length: The prediction of an exact survival in years was the last task and differed from the tasks discussed above as the dependent outcome feature were continuous (Section 6.5). The results showed that models trained in the data did not perform well enough to predict the exact length of survival, and in most cases the predicted values were several years from the known ground truth.

Overall performance of the R^2 score strongly indicated no reliability in performance, and little presence of variation in predicted outcomes were observed from testing ability to generalize. Several regression metrics were used to assess predictions from this task and provided valuable insight into the amount of error and variation in model performance. Looking at the differences between scores based on which features were used indicated the primary surgery reasons features would give a slightly improved result over device features. While increasing the feature space to include both device and primary surgery reasons did not further increase the performance with this collection of records (Section 6.5)

8.2 Explanatory power

The exploration indicated no presence of a correlation between revision indicator and other variables, going into individual clusters and examining record showed the same tendency of practically no correlations. Although, there are larger difference in device usage and materials, the dataset is largely consisting of one similar type of device. The distribution of records according to primary reasons for surgery are also predominantly of one kind. On variables with specific detailed measurements there was a sparsity, with the *caput diameter* as the only one detailing a difference in device specifications. Different types of wear on bone and the device have been suggested to be caused by contrasts in size (Iyer, 2013, p. 328), and more details on device specification could be a suggestion for increasing performance. However, the correlations were not found to be significant between revision indicator and other variables assessed during exploration, presence of causations may still be present by not identifiably in this data.

The modelling in Chapter 6 were presented starting with the binary outcome first, as it has the highest probability of a correct classification with only two possibilities. The highest AUC score achieved was by the multi-layer perceptron classifier at 0.74. The multinomial classification with three classes had a much lower AUC score, suggesting the data could not explain the occurrence of revision surgery to any approvable degree of certainty based on the used features. Investigating results in Chapter 7 showed the distribution of records between classes varied more between predictions and the truth in multinomial classification than observed in binary classification. The AUC were calculated one-vs-the rest, also indicating a disproportionate difference between results on predicted classes (Section 7.2). The accuracy from testing on different sections of the data showed majority of a ten-fold cross validation below %50 in accuracy and stability in performance, suggesting the model might not have learned what it need to make correct decisions on the subject. The last individual outcome prediction was the exact survival year of a prosthetic device until a patient is likely to need revision surgery, which aslo gave a similar result, with larger presence of errors and a stable performance. The results suggest, apart from an above .70 AUC on the binary classification, that this dataset does not contain the adequate information to correctly predict more complex outcomes in number of classes or more precise and detailed in nature (Section 7.2).

The evaluation of results was carried out by metrics provided by the *Scikit-learn* library, as it provides functionality for evaluating results for the machine learning method implemented to explore and predict.

8.3 Challenges and limitations

This section discusses the challenges in development during this research and limitations potentially impacting the results.

Challenges in development

The development in this thesis concerned exploring the data, and training and evaluating predictive models. The larger difficulties experienced were related to exploring the dataset and visualising the data in a convenient and informative way. As well as, dealing with the discovery of new information and having to respond to these changes in an efficient manner. For developing the agile methodology Crystal clear was chosen, which in light of dedaling with the uncovering the unknown, provided an excellent structure for manoeuvring in response to new knowledge and navigating a situation characterized by novice developer experience.

Limitations

The research has a set of limitation which should be considered. The dataset only has a smaller number of revision records left after removing all without revision as a known ground truth, 5538 records in total. There was also observed a skewed distribution according to prosthetic device product types, with one product used in almost all cases. This leaves only a smaller selection of records representing a larger variety in products and materials used in these devices. The records have a larger similarity in reasons for why they required the primary surgery in the first place, with a majority one reason and the others sparsely represented. Overall, there was very little variety in the data.

8.4 Answering research questions

This section addresses the research questions presented in Section 1.2. There were three questions in the focal point of this thesis, and they are answered below.

***Q1**: Which variables in the dataset are suitable as dependent outcome features in this excerpt from a quality registry on hip arthroplasty?*

The dataset had overall only one variable denoting a specific outcome, this being the *revision indicator*. It had as well several descriptive variables detailing more about how and when a revision surgery occurred. The variables describing *length of survival* since a prosthetic

device was implanted was one of the variables detailing a specific measurement describing an outcome if combined with a positive revision indicator (Chapter 5).

Combining these variables gives the opportunity to engineer dependent outcome features to organize records by different classes based on how long the device lasted before revision. In this thesis two dependent features were established by this approach, one dependent outcome variable for binary classification and one for a multinomial classification (Section 5.6.1). The dataset additionally has variables describing why a revision surgery was necessary, suggesting a possibility for combining *revision indicator* and *types of adverse event* (Section 5.6). This could be used for engineering categorical features describing reasons for revision, providing clinicians and patients with the knowledge about what the cause for concern might be in the context of their individual situation.

Q2*: Which variables in the dataset have potential as independent features for explaining an outcome after hip replacement surgery?*

The dataset has a number of variables, but in this dataset no clear correlation between a positive revision indicator and independent features was located. Rather, the opposite was observed, variables available for explaining why a revision occurs indicated that there were no significant association present (Section 5.4.4). Several features were tested for how much they related to the engineered dependent outcome variables and to the exact survival year. However, only a minor selection of variables indicated to have an impact (Section 6.2). These were the *age* of the patient, the *size of the caput* component on the device, and the *material of the caput*. The rest of the variables were indicated to have little or no impact based on the combination of records in the data (Chapter 6).

Q3*: Can the dataset and a selection of learning algorithms give reliable results in predicting an individual patient outcome?*

In answering this question three separate perspectives were taken to test if a reliable result could be achieved in predicting an individual patient outcome. Furthermore, this was done by rotating the use of features to see if any consequential change would be observed by altering some of the independent features. The results showed that the more complex task of classifying records among more than two classes were not reliable, with the multinomial classification producing largely inaccurate results (Section 6.4). Furthermore, it showed to produce a skewed distribution of class membership by classifying a majority of records as belonging to either two of the three classes (Section 7.2). Suggesting the explanatory power in independent feature are not enough for producing a feasible multinomial classification by utilizing this dataset.

Similarly, the preciseness of predicting exact survival length of a prosthetic device before a revision surgery would be required, proved to be an ambitious task based on the available independent features (Section 6.4). The model would deviate with several years from the known truth, the results also showed a larger amount of occurring errors (Section 6.4).

The binary classification was the least complex of the three tasks, with only two possible outcome classes. The results were more positive than what was observed in the tasks discussed above. Looking at performance the model scored an AUC score of above 0.70 (Section 6.3), suggesting the features may have some explanatory power. However, the overall results from modelling indicate that to achieve a more reliable result in predicting an individual patient outcome there is a necessity for supplementary data to convey the underlying relationships to the learning algorithm.

9. Conclusion and future work

This thesis demonstrated the exploration for possible dependent outcome features in a dataset describing hip arthroplasty records. Furthermore, several predictive models were constructed based on the findings of the exploration. The exploration led to the engineering of two outcome features, and a selection of in total three dependent features. Permitting the training and testing of the three separate perspectives on predictive modelling in hip arthroplasty.

Conclusion

Machine learning was performed in two stages, one in exploration and the other in prediction in order to generate models for individual patient outcome in hip arthroplasty. The clinically interesting question is to understand reasons for revisions and predict whether a patient is at risk of a certain outcome. The achieved results showed how the data was structured and how new features could be engineered to suggest new predictive solutions and provide better performing models. In the best case the AUC score was 0.75, which was the *Multi-layer perceptron classifier* for binary classification. All other perspectives on predicting an outcome showed to give results inappropriate for practical use, as the models attempting at classifying with multiple possible outcomes performed unsatisfactory. Even though the combination of methods appears efficient in the binary classification, the limitations of the data could not be overcome in case of increasing complexity of predicted outcome. This would suggest that for better predictions more variety in data could possibly improve performance. There is also the case that more specific details on product specifications could help distinguish between records based on outcomes. This performance is something which supplementary data on a person’s physical status could improve by giving more context about the patient.

Features

The results from modelling showed no clear improvement in predicting with models trained on different features, as each perspective on predicting an outcome was tested three times with variation in independent features in the dataset. There was a slight improvement on using the primary surgery reasons as additional features in comparison to the features describing a prosthetic device. Increasing the feature space to include all independent features did not cause any improvement in the performance of models (Sections 6.3-6.5). This was a reoccurring phenomenon in both the two classification tasks and in the linear regression task.

Learning algorithms
The thesis used in total three machine learning methods for establishing predictive models. These were *Logistic regression*, *Random forest classifier*, and *Multi-layer perceptron*. In both the classification tasks and the prediction of exact survival year, the *Multi-layer perceptron* had the best performance. The other two algorithms performed slightly poorer, though not by much.

Future work

The section briefly discusses the potential for future work following the results achieved in this thesis. There are several aspects which could be beneficial for improving results, including larger quantities of data, as well as more detailed data. Further analysis of correlation between interesting variables within the national registry and investigation of causation for revision surgery could be beneficial. This regards the selection and engineering of more optimal features to train an algorithm to understand the relationships present in hip arthroplasty registry.

There is the interesting activity of piecing together the models and methods into an instantiation, or a full software solution. Making it suitable for performing goal-directed actions within its intended environment and for evaluating the real-world usage of predictive decision aids in hip arthroplasty.

Expansive data analysis
In conducting further research on the subject of building and evaluating individual patient outcome prediction models for hip arthroplasty, a more detailed and expansive data analysis could prove beneficial. Increasing the size of the dataset, not for training a model, but for exploring correlations and possible causations could highlight important constructs in the database.

Supplementary data
Increasing the size of the dataset used for training the algorithm could potentially improve the performance. However, there is not only the aspect of number of records when discussing size of a dataset, as what is described within also matters greatly. In our case the dataset was sparse in variety, with most records belonging to a clear majority in product usage and corresponding device details. As well as, reasons for requiring surgery, and in the availability of detailed measurements on the patient and device. Increasing the data not just by number of records, but in internal contents. i.e. variation in values of different attributes, could prove beneficial.

Learning algorithms

The Scikit-learn framework has a wide variety of learning algorithms, and deciding on the appropriate choice for the task at hand can have an impact on the performance. Future research could as well test and evaluate methods to locate a more optimal choice. Furthermore, there is the area of deep learning concerned with complex neural networks for building prediction models. *The Multi-layer perceptron* in this thesis only consisted of three layers, not taking the full step to deep learning with an increased number of neuron layers. Increasing the complexity of the algorithms might enhance the performance.

Bibliography:

Akoglu, H. (2018) ‘User’s guide to correlation coefficients’, *Turkish Journal of Emergency Medicine*, 18(3), pp. 91–93. doi: 10.1016/j.tjem.2018.08.001.

Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. <https://anaconda.com>.

Buitinck, L. *et al.* (2013) ‘API design for machine learning software: experiences from the scikit-learn project’, pp. 1–15.

Bozic, K. J. *et al.* (2013) ‘Shared decision making in patients with osteoarthritis of the hip and knee results of a randomized controlled trial’, *Journal of Bone and Joint Surgery - Series A*, 95(18), pp. 1633–1639. doi: 10.2106/JBJS.M.00004.

Carreira-Perpiñán, M. Á. (2015) ‘A review of mean-shift algorithms for clustering’, pp. 1–28.

Comaniciu, D. and Meer, P. (2002) ‘Mean shift: A robust approach toward feature space analysis’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), pp. 603–619. doi: 10.1109/34.1000236.

Cabitza, F., Locoro, A. and Banfi, G. (2018) ‘Machine Learning in Orthopedics: A Literature Review’, *Frontiers in Bioengineering and Biotechnology*, 6(June). doi: 10.3389/fbioe.2018.00075.

‘Cockburn, A. (2014) ‘Crystal clear.’, *Nature*, 505(7485), p. 586. doi: 10.1038/505586a.

Calinski, T. and Harabasz, J. (1974) ‘A Dendrite Method for Cluster Analysis’, *Communications in Statistics - Simulation and Computation*, 3(1), pp. 1–27. doi: 10.1080/03610917408548446.

Claesen, M. and De Moor, B. (2015) ‘Hyperparameter Search in Machine Learning’, pp. 10–14.

Cnudde, P. *et al.* (2016) ‘Linking Swedish health data registers to establish a research database and a shared decision-making tool in hip replacement’, *BMC Musculoskeletal Disorders*. BMC Musculoskeletal Disorders, 17(1), p. 414. doi: 10.1186/s12891-016-1262-x.

Delaunay, C. (2014) ‘Registries in orthopaedics’. Orthopedics and Traumatology: Surgery and Research 101. p. 69-75.

Ellison, P., Høl, P. J. and Babic, A. (2018) ‘An Individual Patient Outcome Tool for Joint Replacement Patients’, *Studies in Health Technology and Informatics*, 251, pp. 129–132. doi: 10.3233/978-1-61499-880-8-129.

Fayyad, U. (1994) ‘Data mining and knowledge discovery in databases: implications for scientific databases’, in *Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150)*. IEEE Comput. Soc, pp. 2–11. doi: 10.1109/SSDM.1997.621141.

Fayyad, U. & Uthurusamy, R. (1996) ‘Data Mining and Knowledge Discovery in Databases’. Communications of the ACM. Vol. 39(11). p. 24-26.

Furnes, O., Jan-Erik, G., Hallan, H. Visner, T. Gundersen, I. A. Kvinneland, A. M. Fenstad, E. Dybvik, G. Krokan (2019) ‘Report June 2019’, National Arthroplasty Registry. Haukeland University Hospital.

Fontana, M. A. *et al.* (2019) ‘Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty?’, *Clinical Orthopaedics and Related Research*, 477(6), pp. 1267–1279. doi: 10.1097/CORR.0000000000000687.

Fraley, C. and Raftery, A. E. (2002) ‘Model-based clustering, discriminant analysis, and density estimation’, *Journal of the American Statistical Association*, 97(458), pp. 611–631. doi: 10.1198/016214502760047131.

Glasgow, J., Jurisica, I. and Ng, R. (2000) ‘Data mining and knowledge discovery in molecular databases’, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 39(12), pp. 365–6.

Gorunescu, F. (2012) ‘Data Mining: Concepts, Models, and Techniques’. Intelligent Systems Reference Library, Vol 12. Springer.

Géron, A. (2017) *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. *O'Reilly Media, Sebastopol, CA.*

Gregor, S. and Hevner, A. R. (2013) ‘POSITIONING AND PRESENTING DESIGN SCIENCE Types of Knowledge in Design Science Research’, *MIS Quarterly*, 37(2), pp. 337–355. doi: 10.2753/MIS0742-1222240302.

Hajian-Tilaki, K. (2013) ‘Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation’, *Caspian Journal of Internal Medicine*, 4(2), pp. 627–635.

Hevner *et al.* (2004) ‘Design Science in Information Systems Research’, *MIS Quarterly*, 28(1), p. 75. doi: 10.2307/25148625.

Iyer, K. M. (2013) *General Principles of Orthopedics and Trauma, General Principles of Orthopedics and Trauma*. Edited by K. M. Iyer. London: Springer London. doi: 10.1007/978-1-4471-4444-1.
Kruse, C., Eiken, P. and Vestergaard, P. (2017) ‘Machine Learning Principles Can Improve Hip Fracture Prediction’, *Calcified Tissue International*. Springer US, 100(4), pp. 348–360. doi: 10.1007/s00223-017-0238-7.

Koh, H. C. and Tan, G. (2005) ‘Data mining applications in healthcare.’, *Journal of healthcare information management : JHIM*, 19(2), pp. 64–72.

Konopka, J. F. *et al.* (2015) ‘Risk Assessment Tools Used to Predict Outcomes of Total Hip and Total Knee Arthroplasty’, *Orthopedic Clinics of North America*. Elsevier Inc, 46(3), pp. 351–362. doi: 10.1016/j.ocl.2015.02.004.

Kodinariya, T. M. and Makwana, P. R. (2013) ‘Review on determining number of Cluster in K-Means Clustering’, *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), pp. 2321–7782.

Kluyver, T. *et al.* (2016) ‘Jupyter Notebooks—a publishing format for reproducible computational workflows’, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90. doi: 10.3233/978-1-61499-649-1-87.

Mackinnon, M. J. and Glick, N. (1999) ‘Data Mining and Knowledge Discovery in Databases - An Overview’, *Australian New Zealand Journal of Statistics*, 41(3), pp. 255–275. doi: 10.1111/1467-842X.00081.

March, S. T. and Smith, G. F. (1995) ‘Design and natural science research on information technology’, *Decision Support Systems*, 15(4), pp. 251–266. doi: 10.1016/0167-9236(94)00041-2.

Nelson, E. C. *et al.* (2016) ‘Patient focused registries can improve health, care, and science’, *BMJ*, 354(July), p. i3319. doi: 10.1136/bmj.i3319.

Nemes, S., Rolfson, O. and Garellick, G. (2018) ‘Development and validation of a shared decision-making instrument for health-related quality of life one year after total hip replacement based on quality registries data’, *Journal of Evaluation in Clinical Practice*, 24(1), pp. 13–21. doi: 10.1111/jep.12603.

Obermeyer, Z. and Emanuel, E. J. (2016) ‘Predicting the Future — Big Data, Machine Learning, and Clinical Medicine’, *New England Journal of Medicine*, 375(13), pp. 1216–1219. doi: 10.1056/NEJMp1606181.

Roski, J., Bo-Linn, G. W. and Andrews, T. A. (2014) ‘Creating value in health care through big data: Opportunities and policy implications’, *Health Affairs*, 33(7), pp. 1115–1122. doi: 10.1377/hlthaff.2014.0147.

Rousseeuw, P. J. (1987) ‘Silhouettes: A graphical aid to the interpretation and validation of cluster analysis’, *Journal of Computational and Applied Mathematics*, 20(C), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

Rossum, G van. and F. L. Drake. (2009). *Python 3 Reference Manual*. CreateSpace, Paramount, CA.

Salassa, T. *et al.* (2014) ‘Efficacy of revision surgery for the dislocating total hip arthroplasty: Report from a large community Registry’, *Clinical Orthopaedics and Related Research*, 472(3), pp. 962–967. doi: 10.1007/s11999-013-3344-5.

Sarkar, D., Bali, R. and Sharma, T. (2018) *Practical Machine Learning with Python*. Berkeley, CA: Apress. doi: 10.1007/978-1-4842-3207-1.

Tarasevic˘ius, S. *et al.* (2014) ‘First outcome results after total knee and hip replacement from the Lithuanian arthroplasty register’, *Medicina (Lithuania)*, 0(2), pp. 87–91. doi: 10.1016/j.medici.2014.06.004.

Virtanen, P. *et al.* (2019) ‘SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python’, pp. 1–22.

Varnum, C. *et al.* (2019) ‘Impact of hip arthroplasty registers on orthopaedic practice and perspectives for the future’, *EFORT Open Reviews*, 4(6), pp. 368–376. doi: 10.1302/2058-5241.4.180091.

Varoquaux, G. *et al.* (2015) ‘Scikit-learn’, *GetMobile: Mobile Computing and Communications*, 19(1), pp. 29–33. doi: 10.1145/2786984.2786995.

Yoo, I. *et al.* (2012) ‘Data Mining in Healthcare and Biomedicine: A Survey of the Literature’, *Journal of Medical Systems*, 36(4), pp. 2431–2448. doi: 10.1007/s10916-011-9710-5.

Appendix A

List of variables

This list contains all the variables used either in clustering, and classification and regression, and a review of what values are present in each variable.

1. **REVISJON**

   0 = No Revision.
   1 = Revision.
2. **SURVYRS**

   From 0 and upward = Length of survival since primary operation or revision.
3. **PAS_KJONN**

   1 = Male.
   2 = Female.
4. **P_TIDL_OP**

   0 = Yes.
   1 = No.
   9 = Missing information.
5. **P_ASA**

   0 = Not entered.
   1 = Healthy.
   2 = Mild systemic disease.
   3 = Severe systemic disease.
   4 = Sever systemic disease with treat of death.
   5 = Morbid person, the person is not thought to survive without undergoing surgery.
   9 = Missing information.
6. **P_CUP_MATERIALE**

   0-12 = See list of materials in appendix X, the list of materials coding.
7. **P_LINER_MATERIALE**

   0-12 = See list of materials in appendix X, the list of materials coding.
8. **P_STEM_MATERIALE**

   0-12 = See list of materials in appendix X, the list of materials coding.
9. **P_PROX_MATERIALE**

   0-12 = See list of materials in appendix X, the list of materials coding.
10. **P_DIST_MATERIALE**

    0-12 = See list of materials in appendix X.

11. **P_CAPUT_MATERIALE**

    0-12 = See list of materials in appendix X.

12. **P_CAPUT_DIAMETER:**

    0 = Not entered.
    1.0 and upward = Size of the head on the device.

13. **P_AKT_OP_1-10:**

    Section of dichotomously represented reasons for requiring primary surgery.

Appendix B

List of materials

This list contains all the numbers used to indicate a type of material used in a device and which materials they represent.

0. No materials used or present in the data.
1. UHMWPE.
2. Highly crosslinked polyethylene.
3. Hylamer.
4. Steel.
5. Titanium.
6. Cobolt Chrome.
7. Alumina.
8. Alumina/Zirconium.
9. UHMWPE/Alumina'.
10. Oxinium.
11. Zirconium.
12. Unknown information

Appendix C

Table headers

This list contains all the headers and data recorded in the tables on clustering outcome in Chapter 4, tables are in Appendix D, E, and F.

1. **G1**
    - Number of cases belonging to group 1 from Chapter 4.
2. **G2**
    - Number of cases belonging to group 2 from Chapter 4.
3. **G3**
    - Number of cases belonging to group 3 from Chapter 4 (Gold cases).
4. **M.Age**
    - Mean age for all case in a cluster.
5. **U.Gender**
    - Unique occurrences of gender per cluster.
6. **U.Asa**
    - Unique occurrences of ASA-class per cluster.
7. **Revision**
    - All cases with a revision.
8. **Survyrs**
    - Mean survival year for all cases.
9. **R.Survyrs**
    - Mean survival year for all cases with a revision.
10. **UM.Cup**
    - Unique occurrences of cup materials registered per cluster.
11. **UM.Liner**
    - Unique occurrences of liner materials register per cluster.
12. **UM.Caput**
    - Unique occurrences of caput materials registered per cluster.

13. **UM.Stem**
    - Unique occurrences of stem materials registered per cluster.
14. **UM.Prox**
    - Unique occurrences of prox materials registered per cluster.
15. **UM.Dist**
    - Unique occurrences of dist materials registered per cluster.
16. **M.OPT**
    - Mean surgery time per cluster.
17. **COMP**
    - Number of cases with complications during surgery per cluster.
18. **R.LCUP**
    - Number of revision cases where the cause is a loose cup.
19. **R.LFEMUR**
    - Number of revision cases where the cause is a loose femur.
20. **R.BOTH**
    - Number of revision cases where with both a loose femur and cup.
21. **R.ANNET**
    - Number of revision cases with another reason for reoperation.
22. **R.MANGLER**
    - Number of revision cases where the information is marked as missing/unknown.
23. **P_OP_1**
    - Number of cases requiring primary surgery due to coxarthrosis.
24. **P_OP_2**
    - Number of cases requiring primary surgery due to rheumatoid Arthritis
25. **P_OP_3**
    - Number of cases requiring primary surgery due to sequelae after fracture.
26. **P_OP_4**
    - Number of cases requiring primary surgery due to sequelae after dysplasia.
27. **P_OP_ANNET**
    - Number of cases requiring primary surgery due to another reason.

Appendix D

K-Means below 5 years table results

Groups	G1	G2	G3	M.Ag	U.Kjon	U.Asa	Revisi	Survyrs	R.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET
Group 1	325	0	0	76,87	1.,2.	0.,1.,2.,3.,4.,9.	54	2,571	1,791	0.,1.,2.	0.,1.,2.	6.,7.,8.,10.,11.,0.	0.	1.,2.	6.,7.,8.,10.,11.,0.	97,58	14	11	11	5	307
Group 2	0	1526	0	76,69	1.,2.	0.,1.,2.,3.,4.,9.	271	2,632	2,149	1.	0.	4.,6.,7.,10.,11.,0	0.,6.	1.	4.,6.,7.,10.,11.,0.	104,43	41	87	67	35	1406
Group 3	0	0	7403	75,03	1.,2.	0.,1.,2.,3.,4.,9.	1655	2,656	2,409	0.,1.	0.	0.,4.	0.,4.	0.,1.	0.,4.	111,322	265	348	1025	262	6283

Original groups based on the prduct specifications discussed in Chapter 4, with a survival in years below 5.

Clusters	G1	G2	G3	M.Ag	U.Kjon	U.Asa	Revisi	Survyrs	M.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET	R.MAN
Kmeans G1	322	1510	3	76,73	1.,2.	0.,1.,2.,3.,4.,9.	323	2,619	2,073	2.,1.,0.	0.,1.,2.	4.,6.,7.,8.,10.,11.	0.,6.	1.,2.	4.,6.,7.,8.,10.,11.	103,1	54	96	78	40	1699	2
Kmeans G2	3	16	7400	75,03	1.,2.	0.,1.,2.,3.,4.,9.	1657	2,65	2,41	1.,2.,0.	0.	0.,4.	0.,4.,6.	1.,2.,0.	0.,4.	111,33	266	350	1025	262	6297	9

Summary of results with K = 2 clusters, with survival below 5 years.

Clusters	G1	G2	G3	M.Ag	U.Gend	U.Asa	Revisi	Survyrs	R.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET	R.MAN
Kmeans G1	1	16	7400	75,35	1.,2.	0.,1.,2.,3.,4.,9.	1657	2,657	2,41	0.,1.	0.	0.,4.	0.,4.,6.	0.,1.	0.,4.	111,331	266	350	1025	262	6295	9
Kmeans G2	79	1510	3	76,68	1.,2.	0.,1.,2.,3.,4.,9.	286	2,624	2,153	0.,1.	0.,1.,2.	4.,6.,7.,10.,11	0.,6.	1.,2.	4.,6.,7.,10.,11.	104,082	43	88	73	38	1468	1
Kmeans G3	245	0	0	77,08	1.,2.	0.,1.,2.,3.,4.,9.	37	2,575	1,457	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	97,02	11	8	5	2	233	1

Summary of results with K = 3 clusters, with survival below 5 years.

Clusters	G1	G2	G3	M.Ag	U.Kjon	U.Asa	Revisi	Survyrs	R.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET	R.MAN
Kmeans G1	1	12	5715	76,58	1.,2.	0.,1.,2.,3.,4.,9.	0	2,73	0	0.,1.	0.	0.,4.	0.,4.,6.	1.,0.	0.,4.	110,817	208	0	0	0	0	0
Kmeans G2	79	1501	2	76,75	1.,2.	0.,1.,2.,3.,4.,9.	276	2,63	2,19	0.,1.	0.,1.,2.	4.,6.,7.,10.,11.	0.,6.	1.,2.	4.,6.,7.10.,11.	104,13	43	88	73	38	1458	1
Kmeans G3	0	13	1686	69,74	1.,2.	0.,1.,2.,3.,4.,9.	1667	2,4	2,403	1.,	0.,	4.,6.,0.	0.,4.,6.	1.	4.,6.,0.	112,97	58	350	1025	262	577	9
Kmeans G4	245	0	0	77,08	1.,2.	0.,1.,2.,3.,4.,9.	37	2,575	1,457	2.	0.	6.,7.,8.,10.,0	0.	2.	6.,7.,8.,10.,0	97,02	11	8	5	2	233	1

Summary of results with K = 4 clusters, with survival below 5 years.

Clusters	G1	G2	G3	M.Ag	U.Kjon	U.Asa	Revisi	Survyrs	R.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET
Kmeans G1	0	13	1686	69,74	1.,2.	0.,1.,2.,3.,4.,9.	1667	2,4	2,403	1.,	0.,	4.,6.,0.	0.,4.,6.	1.	4.,6.,0.	112,97	58	350	1025	262	577
Kmeans G2	1	12	5715	76,58	1.,2.,	0.,1.,2.,3.,4.,9.	0	2,73	0	1.,0.	0.	4.,0.	0.,6.	1.,0.	4.,0.	110,81	208	0	0	0	0
Kmeans G3	66	1501	2	76,81	1.,2.,	0.,1.,2.,3.,4.,9.	268	2,635	2,186	1.	0.	4.,6.,7.,10.,11.	0.,6.	1.	4.,6.,7.,10.,11.	104,182	43	87	70	37	1448
Kmeans G4	245	0	0	77,08	1.,2.	0.,1.,2.,3.,4.,9.	37	2,575	1,457	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	97,02	11	8	5	2	233
Kmeans G5	13	0	0	68,64	1.,2.	0.,1.,2.,3.,4.,9.	8	2,55	2,463	0.	1.,2.	6.,10.,11.	0.	1.,2.	6.,7.,8.,10.,0.	98,307	11	1	3	1	10

Summary of results with K = 5 clusters, with survival below 5 years.

Clusters	G1	G2	G3	M.Ag	U.Kjon	U.Asa	Revisi	Survyrs	R.Survy	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMF	R.LCUP	R.LFEMU	R.BOTH	R.ANNET	R.MAN
Kmeans G1	0	12	1660	69,9	1.,2.	0.,1.,2.,3.,4.,9.	1667	2,4	2,403	1.	0.	4.,6.,0.	4.,6.,0.	1.	4.,6.,0.	112,86	56	350	1025	262	550	9
Kmeans G2	66	1501	2	76,817	1.,2.	0.,1.,2.,3.,4.,9.	268	2,623	2,185	1.	0.	4.,6.,7.,10.,11.	0.,6.	1.	4.,6.,7.,10.,11.	104,182	43	87	70	37	1448	1
Kmeans G3	1	10	2044	75,24	1.,2.	0.,1.,2.,3.,4.,9.	0	2,643	0	1.	0.	4.,0.	4.,6.,0.	1.	4.,0.	113,96	0	0	0	0	0	0
Kmeans G4	245	0	0	77,08	1.,2.	0.,1.,2.,3.,4.,9.	37	2,575	1,457	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	97,02	11	8	5	2	233	1
Kmeans G5	13	0	0	68,647	1.,2.	0.,2.	8	2,552	2,463	0.	1.,2.	6.,10.,11.	0.	1.,2.	6.,10.,11.	98,307	0	1	3	1	10	0
Kmeans G6	0	3	3697	77,212	2.	0.,1.,2.,3.,4.,9.	0	2,77	0	1.,0.	0.	0.	4.,0.	1.,0.	4.,0.	109,175	148	0	0	0	0	0

Summary of results with K = 6 clusters, with survival below 5 years.

Appendix E

K-Means below 10 years table results

Groups	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Group 1	1302	0	0	75,2	1.,2.	0.,1.,2.,3.,4.,9.	95	6,769	4,277	1.,2.,0.,	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	0.	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	97,152	52	34	31	17	1252	2
Group 2	0	3967	0	75,9	1.,2.	0.,1.,2.,3.,4.,9.	752	5,74	5,771	1.	0.	4.,6.,7.,10.,11.,0.	0.,5.,6.	1.	4.,6.,7.,10.,11.,0.	103	95	462	265	187	3424	3
Group 3	0	0	17003	74,8	1.,2.	0.,1.,2.,3.,4.,9.	2812	5,429	4,43	1.,0.	0.	0.,4.	0.,4.	1.,0.	4.,0.	108,277	558	700	1876	526	14937	16

Original groups based on the prduct specifications discussed in Chapter 4, with a survival in years below 10.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	1298	3914	6	74,74	1.,2.	0.,1.,2.,3.,4.,9.	840	6	5,602	1.,2.,0.	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	0.,5.,6.	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	101,51	145	490	293	201	4631	5
Kmeans G2	4	53	16997	74,86	1.,2.	0.,1.,2.,3.,4.,9.	2819	5,429	4,437	1.,2.,0.	0.	4.,6.,0.	4.,6.,0.	1.,2.,0.	4.,6.,0.	108,27	560	706	1879	529	14982	16

Summary of results with K = 2 clusters,with survival below 10 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	2	48	16997	74,86	1.,2.	0.,1.,2.,3.,4.,9.	2818	5,431	4,438	0.,1.	0.	0.,4.,6.	0.,4.,6.	1.,0.	4.,6.,0.	108,27	560	706	1879	529	14975	16
Kmeans G2	202	3919	6	75,82	1.,2.	0.,1.,2.,3.,4.,9.	784	5,737	5,735	0.,1.	1.,2.,0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,2.,0.	4.,6.,7.,10.,11.	102,88	98	469	275	189	3568	4
Kmeans G3	1098	0	0	75,409	1.,2.	0.,1.,2.,3.,4.,9.	57	6,954	3,68	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	96,368	47	21	18	12	1070	1

Summary of results with K = 3 clusters,with survival below 10 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	202	3890	5	75,88	1.,2.	0.,1.,2.,3.,4.,9.	754	5,745	5,77	0.,1.	0.,1.,2.	4.,6.,7.,10.,11.	0.,5.,6.	0.,1.,2.	4.,6.,7.,10.,11.	102,97	98	454	270	186	3555	4
Kmeans G2	2	40	14130	76,125	1.,2.	0.,1.,2.,3.,4.,9.	0	5,633	0	0.,1.	0.	0.,4.,6.	0.,4.,6.	0.,1.	0.,4.,6.	107,55	459	0	0	0	0	0
Kmeans G3	0	37	2868	68,64	1.,2.	0.,1.,2.,3.,4.,9.	2848	4,439	4,442	0.,1.	0.	0.,4.,6.	0.,4.,6.	0.,1.	0.,4.,6.	111,6	101	721	1884	532	816	16
Kmeans G4	1098	0	0	75,4	1.,2.	0.,1.,2.,3.,4.,9.	57	6,95	3,68	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	96,368	47	21	18	12	1070	1

Summary of results with K = 4 clusters, with survival below 10 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	2	40	14130	76,12	1.,2.	0.,1.,2.,3.,4.,9.	0	5,632	0	0.,1.	0.	0.,4.,6.	0.,4.,6.	0.,1.	0.,4.,6.	107,55	459	0	0	0	0	0
Kmeans G2	1098	0	0	75,4	1.,2.	0.,1.,2.,3.,4.,9.	57	6,95	3,68	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	96,368	47	21	18	12	1070	1
Kmeans G3	44	0	0	65,69	1.,2.	1.,2.,3.,0.,9.	18	6,36	5,171	0.	1.,2.	6.,10.,11.	0.	1.,2.	6.,10.,11.	101,477	0	3	9	2	33	1
Kmeans G4	158	3890	5	75,99	1.,2.	0.,1.,2.,3.,4.,9.	736	5,738	5,789	0.,1.	0.	4.,6.,7.,10.,11.	0.,5.,6.	0.,1.	4.,6.,7.,10.,11.	102,98	98	451	261	184	3522	3
Kmeans G5	0	38	2838	68,588	1.,2.	0.,1.,2.,3.,4.,9.	2849	4,43	4,44	0.,1.	0.	0.,4.,6.	0.,4.,6.	0.,1.	0.,4.,6.	111,58	101	721	1884	532	787	16

Summary of results with K = 5 clusters,with survival below 10 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	1	20	9634	76,78	2.	0.,1.,2.,3.,4.,9.	0	5,77	0	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	106,46	331	0	0	0	0	0
Kmeans G2	158	3889	5	75,98	1.,2.	0.,1.,2.,3.,4.,9.	735	5,739	5,797	1.,0.	0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,0.	4.,6.,7.,10.,11.	102,96	98	451	261	184	3521	3
Kmeans G3	1098	0	0	75,4	1.,2.	0.,1.,2.,3.,4.,9.	57	6,954	3,68	2.	0.	6.,7.,8.,10.,0.	0.	2.	6.,7.,8.,10.,0.	96,36	47	21	18	12	1070	1
Kmeans G4	0	37	2819	68,75	1.,2.	0.,1.,2.,3.,4.,9.	2849	4,43	4,44	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	111,44	96	721	1884	532	767	16
Kmeans G6	1	21	4545	74,57	1.,2.	0.,1.,2.,3.,4.,9.	0	5,32	0	1.,0.	0.	4.,0.	4.,6.,0.	1.,0.	4.,6.,0.	109,93	133	0	0	0	0	0
Kmeans G5	44	0	0	65,69	1.,2.	0.,1.,2.,3.,9.	18	6,362	5,171	0.	1.,2.	6.,10.,11.	0.	1.,2.	6.,10.,11.	101,47	0	3	9	2	33	1

Summary of results with K = 6 clusters with survival below 10 years.

Appendix F

K-Means below 15 years table results

Groups	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Group 1	2162	0	0	74,1	1.,2.	0.,1.,2.,3.,4.,9.	118	8,66	5,75	1.,2.,3,.0.	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	0.	1.,2.,3.,0.	4.,6.,7.,8.,10.,11.,4	95,366	63	45	40	19	2094	2
Group 2	0	7824	0	73,96	1.,2.	0.,1.,2.,3.,4.,9.	1226	9,07	8,23	1.	0.	4.,6.,7.,10.,11.,0.	0.,5.,6.	1.	4.,6.,7.,10.,11.,0.	100,794	152	864	429	330	6855	6
Group 3	0	0	27420	73,85	1.,2.	0.,1.,2.,3.,4.,9.	3503	8,123	5,99	1.,0	0.	4.,7.,0.	4.,0.	1.,0.	4.,7.,0.	107,297	784	976	2311	715	24828	20

Original groups based on the product specifications discussed in Chapter 4, with a survival in years below 15.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	10	64	27406	73,85	1.,2.	0.,1.,2.,3.,4.,9.	3512	8,121	5,99	1.,2.,0.	1.,0.	4.,6.,0.	4.,6.,0.	1.,2.,0.	4.,6.,0	107,28	786	984	2316	719	24879	20
Kmeans G2	2152	7760	14	73,98	1.,2.	0.,1.,2.,3.,4.,9.	1335	8,99	8,01	1.,2.,3.,0.	1.,2.,0.	4.,6.,7.,8.,10.,11.,0.	5.,6.,0.	1.,2.,3.,0.	4.,6.,7.,8.,10.,11.,0.	99,59	213	901	464	345	8898	8

Summary of results with K = 2 clusters,with survival below 15 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	M.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	3	63	27406	73,85	1.,2.	0.,1.,2.,3.,4.,9.	3512	8,121	5,99	1.,0.	1.,0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	107,29	786	984	2316	719	24871	20
Kmeans G2	429	7761	14	73,89	1.,2.	0.,1.,2.,3.,4.,9.	1278	9,107	8,208	1.,0.	1.,2.,0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,2.,0.	4.,6.,7.,10.,11.	100,93	159	880	446	333	7204	7
Kmeans G3	1730	0	0	74,42	1.,2.	0.,1.,2.,3.,4.,9.	57	8,465	3,68	2.,3.	0.	6.,7.,8.,10.,0.	0.	2.,3.	6.,7.,8.,10.,0.	93,23	54	21	18	12	1702	1

Summary of results with K = 3 clusters,with survival below 15 years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	M.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	0	78	3507	68,02	1.,2.	0.,1.,2.,3.,4.,9.	3581	6,049	6,05	1.,0.	0.	4.,6.,7.,0.	4.,5.,6.,0.	1.,0.	4.,6.,7.,0.	112,13	119	1026	2329	728	939	19
Kmeans G2	429	7693	13	73,94	1.,2.	0.,1.,2.,3.,4.,9.	1209	9,11	8,17	1.,0.	1.,2.,0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,2.,0.	4.,6.,7.,10.,11.,	100,8	158	838	433	324	7181	7
Kmeans G3	1730	0	0	74,42	1.,2.	0.,1.,2.,3.,4.,9.	57	8,465	3,68	2.,3.	0.	6.,7.,8.,10.,0.	0.	2.,3.	6.,7.,8.,10.,0.	93,23	54	21	18	12	1702	1
Kmeans G4	3	53	23900	74,71	1.,2.	0.,1.,2.,3.,4.,9.	0	8,433	0	1.,0.	1.,0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	106,58	668	0	0	0	0	0

Summary of results with K = 4 clusters,with survival below 15 years

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G1	1	29	14916	72,97	1.,2.	0.,1.,2.,3.,4.,9.	546	11,27	12,617	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	104,94	344	22	331	144	14533	4
Kmeans G2	1730	0	0	74,42	1.,2.	0.,1.,2.,3.,4.,9.	57	8,465	3,68	2.,3.	0.	6.,7.,8.,10.,0.	0.	2.,3.	6.,7.,8.,10.,0.	93,239	54	21	18	12	1702	1
Kmeans G3	360	7721	14	73,94	1.,2.	0.,1.,2.,3.,4.,9.	1240	9,15	8,311	1.,0.	0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,0.	4.,6.,7.,10.,11.	100,82	159	877	431	331	7112	6
Kmeans G4	1	74	12490	75,13	1.,2.	0.,1.,2.,3.,4.,9.	2979	4,34	4,76	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	110,03	422	762	1986	575	10376	16
Kmeans G5	70	0	0	64,7	1.,2.	0.,1.,2.,3.,9.	25	8,71	6,98	0.	1.,2.	6.,10.,11.,0.	0.	1.,2.	6.,10.,11.,0.	118,98	0	3	14	2	54	1

Summary of results with K = 5 clusters,with survival below 15 years

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
Kmeans G2	360	7646	13	74,01	1.,2.	0.,1.,2.,3.,4.,9.	1184	9,14	8,2	1.,0.	0.	4.,6.,7.,10.,11.	5.,6.,0.	1.,0.	4.,6.,7.,10.,11.	100,77	158	835	419	322	7058	6
Kmeans G1	1	71	7175	73,26	1.,2.	0.,1.,2.,3.,4.,9.	0	7,921	0	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	108,56	194	0	0	0	0	0
Kmeans G4	1	29	16728	75,33	2.	0.,1.,2.,3.,4.,9.	0	8,642	0	1.,0.	0.	4.,6.,0.	4.,6.,0.	1.,0.	4.,6.,0.	105,67	475	0	0	0	0	0
Kmeans G6	0	78	3504	68,053	1.,2.	0.,1.,2.,3.,4.,9.	3581	6,05	6,05	1.,0.	0.	4.,6.,7.,0.	4.,5.,6.,0.	1.,0.	4.,6.,7.,0.	112,11	118	1026	2329	728	936	19
Kmeans G5	70	0	0	64,74	1.,2.	0.,1.,2.,3.,9.	25	8,713	6,987	0.	1.,2.	6.,10.,11.,0.	0.	1.,2.	6.,10.,11.,0.	118,98	0	3	15	2	54	1
Kmeans G3	1730	0	0	74,42	1.,2.	0.,1.,2.,3.,4.,9.	57	8,46	3,68	2.,3.	0.	6.,7.,8.,10.,0.	0.	2.,3.	6.,7.,8.,10.,11.	93,23	54	21	18	12	1702	1

Summary of results with K = 6 clusters,with survival below 15 years

Appendix G

K-Means below 5 years table results, primary surgery reasons

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	1236	37	394	45	134
KMEANS G2	4713	276	1717	217	504

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	4713	276	1717	217	502
KMEANS G2	1046	35	365	40	111
KMEANS G3	190	2	29	5	25

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	3456	225	1515	155	384
KMEANS G2	1038	25	365	40	109
KMEANS G3	1265	51	202	62	120
KMEANS G4	190	2	29	5	25

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	1265	51	202	62	120
KMEANS G2	3456	225	1515	155	384
KMEANS G3	1026	34	365	40	109
KMEANS G4	190	2	29	5	25
KMEANS G5	12	1	0	0	0

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	1253	50	200	62	108
KMEANS G2	1026	34	365	40	109
KMEANS G3	1415	60	369	58	156
KMEANS G4	190	2	29	5	25
KMEANS G5	12	1	0	0	0
KMEANS G6	2053	166	1148	97	240

Appendix H

K-Means below 10 years, table results for primary surgery reasons

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	3964	89	729	198	294
KMEANS G2	11676	634	3271	510	989

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	11674	634	3268	510	987
KMEANS G2	3034	75	654	147	231
KMEANS G3	932	14	78	51	65

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	3010	75	651	146	229
KMEANS G2	9532	538	2961	364	801
KMEANS G3	2166	96	310	147	188
KMEANS G4	932	14	78	51	65

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	9532	538	2961	364	801
KMEANS G2	932	14	78	51	65
KMEANS G3	36	2	3	2	1
KMEANS G4	2974	73	648	144	228
KMEANS G5	2146	96	307	147	182

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	6155	408	2345	245	519
KMEANS G2	2974	73	648	144	227
KMEANS G3	932	14	78	51	65
KMEANS G4	2141	94	306	147	170
KMEANS G5	3402	132	620	119	301
KMEANS G6	36	2	3	2	1

Appendix I

K-Means below 15 years, table results for primary surgery reasons

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	19935	971	4264	920	1447
KMEANS G2	7962	163	1027	417	449

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	19931	971	4263	920	1444
KMEANS G2	6469	145	925	345	361
KMEANS G3	1497	18	103	72	91

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	2685	121	356	214	212
KMEANS G2	6412	145	920	342	357
KMEANS G3	1497	18	103	72	91
KMEANS G4	17303	850	3912	709	1236

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	11841	552	1350	547	699
KMEANS G2	1497	18	103	72	91
KMEANS G3	6398	141	902	339	356
KMEANS G4	8106	419	2933	373	748
KMEANS G5	55	2	4	6	2

CLUSTERS	P_OP_1	P_OP_2	P_OP_3	P_OP_4	P_OP_ANNET
KMEANS G1	6325	141	909	335	354
KMEANS G2	5626	174	799	200	465
KMEANS G3	11713	67	3121	510	775
KMEANS G4	2685	121	356	214	209
KMEANS G5	55	4	3	6	2
KMEANS G6	1497	18	103	72	91

Appendix J

Mean Shift below 5 and 10 years, table of results

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
1	0	1	7336	75	1.,2.	0.,1.,2.,3.,4.	1643	2,65	2,41	1.	0.	4.,0.	6.,4.	1.	4.	111,4	260	346	1022	261	6222	8
2	72	1429	3	76,8	1.,2.	0.,1.,2.,3.,4.	266	2,61	2,11	1.,0.	0.,1.,2.	4.,6.,7.	6.,0.	1.,2.	4.,6.,7.	103,7	40	79	66	36	1394	1
3	242	0	0	77,02	1.,2.	0.,1.,2.,3.,4.	37	2,59	1,457	2.	0.	6.,7.,8.,10.	0.	2.	6.,7.,8.,10.	96,55	11	8	5	2	230	1
4	1	51	0	79,2	1.,2.	9.	5	2,71	2,64	1.	0.	6.,7.	0.	1.	6.,7.	99,26	2	2	3	1	48	0
5	0	0	35	76,16	1.,2.	9.	5	2,44	1,66	1.	0.	0.	4.	1.	4.	85,88	3	0	0	0	35	0
6	6	28	0	64,87	1.,2.	0.,2.	15	2,7	2,72	1.,0.	0.,1.	10.,11.	0.	1.	10.,11.	127	1	7	4	1	24	0
7	0	0	19	75,16	1.,2.	0.,1.	4	2,95	2,86	1.	0.	0.	4.	1.	4.	119,5	1	1	2	0	16	0
8	1	12	4	78,22	1.,2.	0.,2.	2	2,93	2,86	1.	0.	0.	0.	1.	0.	114,9	1	1	0	0	16	0
9	0	0	3	69,87	2.	0.	0	1,74	0	0.	0.	0.	4.	0.	4.	131,7	0	0	0	0	3	0
10	1	2	0	85,46	2.	2.,3.,0.	0	2,51	0	1.,2.	0.	6.	0.	1.,2.	6.	117,7	0	0	0	0	3	0
11	0	1	0	33,17	1.	0.	0	4,29	0	1.	0.	0.	0.	1.	0.	120	0	0	0	0	1	0
12	0	1	3	29,39	1.,2.	0.	3	3,16	3,12	1.	0.	0.	0.,4.	1.	0.,4.	176,3	1	2	1	1	1	1
13	0	1	0	81,73	2.	9.	0	2,26	0	1.	0.	0.	0.	1.	0.	0	0	0	0	1	0	0
14	2	0	0	78,68	2.	3.,9,	0.	1,28	0	2.	0.	0.	0.	2.	0.	135	0	0	0	0	2	0

Summary of results withMean Shift cluster model and data at 5 years or below in survival years.

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
1	0	5	16851	74,85	1.,2.	0.,1.,2.,3.,4.	2798	5,42	4,43	1.	0.	4.,6.,0.	4.,6.	1.	4.,6.	108,3	546	698	1872	525	14795	16
2	176	3703	6	76,17	1.,2.	0.,1.,2.,3.,4.	696	5,73	5,6	1.,0.	1.,2.,0.	4.,6.,7.	5.,6.,0.	1.,2.	4.,6.,7.	102	91	404	232	159	3405	3
3	1071	0	0	75,4	1.,2.	0.,1.,2.,3.,4.	56	6,98	3,7	2.	0.	6.,7.,8.,10.	0.	2.	6.,7.,8.,10	96,12	44	21	18	12	1043	1
4	5	160	0	78,3	1.,2.	9.	13	5,17	5,7	1.,0.	2.,0.	6.,7.	0.	1.,2.	6.,7.	101,4	5	7	7	3	100	0
5	19	108	0	62,86	1.,2.	0.,1.,2.,3.	75	6,42	6,78	1.,0.	1.,2.,0.	7.,10.,11.	0.	1.,2.	7.,10.,11.	130,2	2	58	36	27	59	1
6	0	0	94	76,55	1.,2.	9.	8	5,76	3,61	1.	0.	0.	4.	1.	4.	95,31	7	0	1	0	93	0
7	2	41	7	74,6	1.,2.	0.,2.,3.,4.	6	5,72	5,2	1.	0.	0.	0.	1.	0.	112,8	3	5	3	3	45	0
8	2	2	40	75,14	1.,2.	0.,1.,2.,3.	4	5,07	2,86	1.,2.	0.	6.,0.	4.,0.	1.,2.	4.,6.	111,9	4	1	2	0	41	0
9	22	0	0	74,9	1.,2.	9.	1	6,49	0,08	2.	0.	6.,7.,10.	0.	2.	6.,7.,10.	102,9	3	0	0	0	22	0
10	1	0	5	72,07	1.,2.	0.,2.,3.	1	4,92	9,81	0.	0.	6.,0.	4.,0.	0.	4.,6.	135,8	0	1	1	1	5	0
11	3	0	0	80,03	1.,2.	1.,2.,3.	0	4,96	0	2.	0.	0.	0.	2.	0.	111,7	0	0	0	0	3	0
12	0	1	0	67,28	2.	9.	1	9	9	1.	0.	0.	0.	1.	0.	165	0	1	0	0	0	0
13	0	1	0	81,73	2.	9.	0	2,26	0	1.	0.	0.	0.	1.	0.	0	0	0	0	0	1	0
14	1	0	0	67,98	2.	9.	0	1,69	0	2.	0.	0.	0.	2.	0.	130	0	0	0	0	1	0

Summary of results withMean Shift cluster model and data at 10 years or below in survival years.

Appendix K

Mean Shift below 15 years table of results

Clusters	G1	G2	G3	M.Age	U.Kjonn	U.Asa	Revisions	Survyrs	R.Survyrs	UM.Cup	UM.Liner	UM.Caput	UM.Stem	UM.Prox	UM.Dist	M.OPT	COMP	R.LCUP	R.LFEMUR	R.BOTH	R.ANNET	R.MANG
1	0	6	27107	73,86	1.,2.	0.,1.,2.,3.,4.	3484	8,1	6	1.	0.	4.,6.,0.	4.,6.	1.	4.,6.	107,4	766	971	2304	711	24529	20
2	376	7285	14	74,21	1.,2.	0.,1.,2.,3.,4.	1117	9,08	8,07	1.,0.	1.,2.,0.	4.,6.,7.	5.,6.,0.	1.,2.	4.,6.,7.	100	150	72	371	277	6824	5
3	1674	0	0	74,4	1.,2.	0.,1.,2.,3.,4.	56	8,4	3,7	2.,3.	0.	6.,7.,8.,10.	0.	2.,3.	6.,7.,8.,10.	93,17	50	21	18	12	1646	1
4	10	242	0	74,33	1.,2.	9.	22	9,2	8	1.,0.	2.,0.	6.,7.,10.	5.,0.	1.,2.	6.,7.,10.	98,99	7	15	8	4	232	1
5	40	231	0	64,31	1.,2.	0.,1.,2.,3.	139	9,61	9,33	1.,0.	1.,2.,0.	7.,10.,11.	0.	1.,2.,0.	7.,10.,11.	127,4	2	113	67	52	142	1
6	0	0	211	72,48	1.,2.	9.	10	9,83	5,24	1.	0.	0.	4.	1.	4.	94,74	11	1	2	1	209	0
7	0	0	70	72,84	1.,2.	0.,1.,2.,3.	5	8,2	5	1.	0.	0.	4.	1.	4.	108,9	5	2	3	1	66	0
8	3	54	12	74,44	1.,2.	0.,1.,2.,3.,4.	11	7,65	8,63	1.,0.	1.,0.	0.	0.	1.	0.	109,5	4	8	6	5	60	0
9	43	0	0	74,87	1.,2.	9.	1	8,8	0,08	2.	0.	6.,7.,10.	0.	2.	6.,7.,10.	92,2	4	0	0	0	43	0
10	10	0	0	75,45	1.,2.	1.,2.,3.	0	10,29	0	2.	0.	0.	0.	2.	0.	99	0	0	0	0	10	0
11	2	0	5	73,35	1.,2.	2.,3.,0.	0	6,35	0	0.	0.	6.,0.	4.,0.	0.	4.,6.	119,3	0	0	0	0	7	0
12	2	3	0	81,3	2.	0.,1.,2.,3.	0	5,68	0	1.,2.	0.	6.	0.	1.,2.	6.	102,6	0	0	0	0	5	0
13	0	2	0	71,86	2.	9.	1	11,08	9,06	1.	0.	0.	0.	1.	0.	130,5	0	1	0	0	1	0
14	0	0	1	70,37	1.	0.	1	9,81	9,81	0.	0.	0.	4.	0.	4.	135	0	1	1	1	0	0
15	0	1	0	81,73	2.	9.	0	2,26	0	1.	0.	0.	0.	1.	0.	0	0	0	0	0	1	0
16	2	0	0	71,19	2.	9.	0	7,9	0	2.	0.	0.	0.	2.	0.	125	0	0	0	0	2	0

Summary of results withMean Shift cluster model and data at 15 years or below in survival years.

Appendix L

Silhouette Coefficient table, Mean Shift

BANDWIDTH	5 YEARS	10 YEARS	15 YEARS
0.1	*0.299*	*0.339*	*0.373*
0.125	*0.376*	*0.353*	*0.368*
0.15	*0.413*	*0.379*	*0.374*
0.175	*0.413*	*0.382*	*0.379*
0.2	*0.414*	*0.430*	*0.415*
0.225	*0.414*	*0.433*	*0.434*
0.25	*0.414*	*0.433*	*0.433*
0.275	*0.414*	*0.447*	*0.432*
0.3	*0.456*	*0.446*	*0.432*
0.325	*0.456*	*0.441*	*0.432*
0.35	*0.458*	*0.441*	*0.432*
0.375	*0.428*	*0.441*	*0.431*
0.4	*0.428*	*0.438*	*0.432*
0.425	*0.428*	*0.437*	*0.432*
0.45	*0.442*	*0.438*	*0.434*
0.475	*0.442*	*0.437*	*0.357*
0.5	*0.442*	*0.442*	*0.357*

Appendix M

NSD Approval

NSD NORSK SENTER FOR FORSKNINGSDATA

NSD sin vurdering

Prosjekttittel

Decision-aid tool for arthroplasty

Referansenummer

159469

Registrert

24.04.2019 av Yngve Kristoffersen - Yngve.Kristoffersen@student.uib.no

Behandlingsansvarlig institusjon

Universitetet i Bergen / Det samfunnsvitenskapelige fakultet / Institutt for informasjons- og medievitenskap

Prosjektansvarlig (vitenskapelig ansatt/veileder eller stipendiat)

Ankica, Ankica.babic@uib.no, tlf: 4755589139

Type prosjekt

Studentprosjekt, masterstudium

Kontaktinformasjon, student

Yngve Kristoffersen, y.kristoffersen@outlook.com, tlf: 41650562

Prosjektperiode

01.06.2019 - 01.12.2019

Status

04.11.2019 - Vurdert

Vurdering (2)

04.11.2019 - Vurdert

Vi viser til endring registrert 18.10.2019. Vi kan ikke se at det er gjort noen oppdateringer i meldeskjemaet eller vedlegg som har innvirkning på NSD sin vurdering av hvordan personopplysninger behandles i prosjektet.

Les mer om hvilke endringer som skal registreres før endringer meldes inn i fremtiden:
http://www.nsd.uib.no/personvernombud/meld_prosjekt/meld_endringer.html

OPPFØLGING AV PROSJEKTET
NSD vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.
Lykke til videre med prosjektet!

Tlf. Personverntjenester: 55 58 21 17 (tast 1)

29.04.2019 - Vurdert

Det er vår vurdering at behandlingen av personopplysninger i prosjektet vil være i samsvar med personvernlovgivningen så fremt den gjennomføres i tråd med det som er dokumentert i meldeskjemaet med vedlegg den 29.04.2019. Behandlingen kan starte.

MELD VESENTLIGE ENDRINGER
Dersom det skjer vesentlige endringer i behandlingen av personopplysninger, kan det være nødvendig å melde dette til NSD ved å oppdatere meldeskjemaet. Før du melder inn en endring, oppfordrer vi deg til å lese om hvilke type endringer det er nødvendig å melde:

https://nsd.no/personvernombud/meld_prosjekt/meld_endringer.html

Du må vente på svar fra NSD før endringen gjennomføres.

TYPE OPPLYSNINGER OG VARIGHET
Prosjektet vil behandle alminnelige kategorier av personopplysninger frem til 01.12.2019.

LOVLIG GRUNNLAG
Prosjektet vil innhente samtykke fra de registrerte til behandlingen av personopplysninger. Vår vurdering er at prosjektet legger opp til et samtykke i samsvar med kravene i art. 4 og 7, ved at det er en frivillig, spesifikk, informert og utvetydig bekreftelse som kan dokumenteres, og som den registrerte kan trekke tilbake. Lovlig grunnlag for behandlingen vil dermed være den registrertes samtykke, jf. personvernforordningen art. 6 nr. 1 bokstav a.

PERSONVERNPRINSIPPER
NSD vurderer at den planlagte behandlingen av personopplysninger vil følge prinsippene i personvernforordningen om:

- lovlighet, rettferdighet og åpenhet (art. 5.1 a), ved at de registrerte får tilfredsstillende informasjon om og samtykker til behandlingen
- formålsbegrensning (art. 5.1 b), ved at personopplysninger samles inn for spesifikke, uttrykkelig angitte og berettigede formål, og ikke behandles til nye, uforenlige formål
- dataminimering (art. 5.1 c), ved at det kun behandles opplysninger som er adekvate, relevante og nødvendige for formålet med prosjektet
- lagringsbegrensning (art. 5.1 e), ved at personopplysningene ikke lagres lengre enn nødvendig for å oppfylle formålet

DE REGISTRERTES RETTIGHETER
Så lenge de registrerte kan identifiseres i datamaterialet vil de ha følgende rettigheter: åpenhet (art. 12), informasjon (art. 13), innsyn (art. 15), retting (art. 16), sletting (art. 17), begrensning (art. 18), underretning

(art. 19), dataportabilitet (art. 20).

NSD vurderer at informasjonen om behandlingen som de registrerte vil motta oppfyller lovens krav til form og innhold, jf. art. 12.1 og art. 13.

Vi minner om at hvis en registrert tar kontakt om sine rettigheter, har behandlingsansvarlig institusjon plikt til å svare innen en måned.

FØLG DIN INSTITUSJONS RETNINGSLINJER
NSD legger til grunn at behandlingen oppfyller kravene i personvernforordningen om riktighet (art. 5.1 d), integritet og konfidensialitet (art. 5.1. f) og sikkerhet (art. 32).

For å forsikre dere om at kravene oppfylles, må dere følge interne retningslinjer og/eller rådføre dere med behandlingsansvarlig institusjon.

OPPFØLGING AV PROSJEKTET
NSD vil følge opp ved planlagt avslutning for å avklare om behandlingen av personopplysningene er avsluttet.

Lykke til med prosjektet!

Kontaktperson hos NSD: Karin Lillevold
Tlf. Personverntjenester: 55 58 21 17 (tast 1)